TEST AUTOKORELACYJNY DLA CIĄGU BINARNEGO

Krzysztof Mańk

Wojskowa Akademia Techniczna, Wydział Cybernetyki,
Laboratorium Badawcze Kryptologii
00-908 Warszawa, ul. S. Kaliskiego 2,
kmank@wat.edu.pl

Streszczenie. W pracy tej prezentujemy wariant testu autokorelacyjnego dedykowany dla ciągów binarnych, dzielonych na bloki bitów. Dla proponowanego testu wyznaczyliśmy przybliżenie rozkładu statystyki testowej oraz przeprowadziliśmy analizę jego jakości. Pokazaliśmy również konsekwencje użycia testu przy podziale ciągu na wielobitowe bloki.

Słowa kluczowe: test statystyczny, test losowości, test autokorelacyjny

Wstęp

W literaturze znaleźć można dwa warianty testu badającego ciąg liczb rzeczywistych zprzedziału $[0, 1)$ pod kątem jego autokorelacji. Pierwszy zaproponowany został przez Donalda Knutha w powszechnie znanej *Sztuce programowania* [5], test ten polega na obliczeniu współczynnika korelacji pomiędzy badanym ciągiem, a jego cyklicznym przesunięciem o t pozycji. Drugi zwariantów znaleźć można, na przykład, wbibliotece *TestU01* [2], przewiduje on jednak wyznaczanie współczynnika korelacji zniecyklicznym przesunięciem ciągu. Dopełnieniem dla nich jest test autokorelacyjny dla ciągu binarnego.

W niniejszej pracy przedstawimy krótko obydwa warianty testu dla ciągu liczb rzeczywistych, skupiając się na poprawności określenia ich parametrów oraz podatności na implementację w środowisku o ograniczonych zasobach. W drugiej części zaproponujemy wariant testu i jego implementacji pozwalający na możliwie kompaktową implementację.

Praca dofinansowana ze środków NCBiR na lata 2013-2018, w ramach projektu numer DOBR/0001/RON/ID1/2013/01.

1. Testy autokorelacyjne dla ciągu liczb rzeczywistych

W tym rozdziale zakładać będziemy, że testowany jest ciąg $U^l = u_1, u_2, \ldots, u_l$ – ciąg l liczb rzeczywistych, zaś u_i tworzą ciąg niezależnych zmiennych losowych o rozkładzie równomiernym na przedziale $[0, 1)$ – co jest własnością weryfikowaną przez testy.

1.1. Test zaproponowany przez Knutha

Knuth zaproponował wyznaczanie statystyki testowej postaci:

$$S_{cykl} = \frac{n\sum_{i=1}^{n} u_i u_{i+t} - \left(\sum_{i=1}^{n} u_i\right)^2}{n\sum_{i=1}^{n} u_i^2 - \left(\sum_{i=1}^{n} u_i\right)^2},$$

przy czym należy przyjąć, że badany ciąg jest postaci $U^{n+t} = u_1, u_2, \ldots, u_n, u_1, \ldots, u_t$.
Dla dużych n można przyjąć, że opisuje ją zmienna losowa o rozkładzie normalnym, z wartością oczekiwaną

$$\mu = -1/(n-1)$$

i wariancją

$$\sigma^2 \approx \frac{n^2}{(n-1)^2(n-2)} - \frac{24}{5n^2}.$$

Przeprowadzone przez nas testy wskazały, że rozkład i parametry statystyki testowej zostały poprawnie określone.

Analiza testu pod kątem implementacji w strukturach programowalnych wskazała na dwie, dość istotne, wady, podnoszące zajętość struktury. Pierwszą z nich jest cykliczność stosowanego przesunięcia, oznacza to konieczność zapamiętania $t-1$ początkowych elementów ciągu, w celu ich późniejszego użycia, drugą – występowanie aż trzech sum w statystyce testowej. Należy zaznaczyć, że obie te niedogodności można stosunkowo prosto przezwyciężyć. W pierwszym przypadku, poprzez przerzucenie ciężaru przechowania lub dwukrotnego wytworzenia wspomnianego fragmentu na stronę dostarczającą badany ciąg, co jednak może być z różnych względów nieakceptowalne, np. gdy układ wytwarzający implementowany jest wewnątrz tej samej struktury. W drugim przypadku rozwiązanie jest już proste i eleganckie, polega na skojarzeniu tego testu z testami momentów pierwszego i drugiego rzędu, na użytek których wyznacza się sumę i sumę kwadratów elementów ciągu.

1.2. Test z niecyklicznym przesunięciem

Implementację testu w przedstawionej poniżej wersji znaleźć można w bibliotece *svaria*, wchodzącej w skład pakietu testów *TestU01*. Występuje tam pod nazwą *sample correlation test*.

Wyznaczana jest statystyka testowa:

$$S_{ncykl} = \frac{1}{n}\sum_{i=1}^{n} u_i u_{i+t} - \frac{1}{4},$$

przy czym, odmiennie niż w poprzednim punkcie, badany jest, po prostu, ciąg $U^{n+t} = u_1, u_2, \ldots, u_{n+t}$.

Podano, że dla dużych n zmienna losowa $S_{ncykl}\sqrt{12n}$ ma w przybliżeniu rozkład $N(0,1)$. Proponuje się przeprowadzanie testu zgodności, jako testu drugiego poziomu, dla wielokrotnych powtórzeń testu przedstawionego.

Poprzez rezygnację z cykliczności przesunięcia udało się usunąć obie niedogodności, które sygnalizowaliśmy w poprzednim punkcie, niestety stronę analityczną autorzy potraktowali po macoszemu. Niezbyt złożone rachunki, pozwalają wyznaczyć prawidłową formułę opisującą wariancję statystyki testowej, dla której otrzymaliśmy:

$$D^2\left[S_{ncykl}\right] = \frac{13}{144n} - \frac{t}{24n^2}.$$

Drugi składnik jest w praktyce zaniedbywalny, wobec czego okazuje się, że rzeczywista wariancja jest o nieco ponad 8% większa od pierwotnie postulowanej. Jakie to może mieć znaczenie pokażemy na poniższym przykładzie.

Załóżmy, że badanie zostało zaprojektowane przy użyciu pierwotnej formuły, na podstawie której wyznaczono kwantyle dziesiętne – test drugiego poziomu będzie testem zgodności Poissona, w którym rozróżnianych jest 10 klas, zaś poddany mu został generator posiadający założone własności. Tabela 1 przedstawia wartości pierwszych 5 kwantyli – pozostałe pominęliśmy ze względu na symetryczność rozkładu – oraz odpowiadających im prawdopodobieństw uwzględniających poprawną wartość wariancji (nie uwzględnialiśmy drugiego składnika).

Takie ustalenie progów miało na celu uzyskanie równego 10% prawdopodobieństwa dla każdej z 10 klas. Jak widać, największe różnice występują dla skrajnych klas, dla których przekraczają one 9% wartości założonej, a spośród pozostałych największe są dla klas środkowych – tu odchylenie

wynosi niecałe 4%.

Tabela 1

Kwantyle dziesiętne i odpowiadające im rzeczywiste prawdopodobieństwa

założone prawdopodobieństwo	kwantyl $\cdot \sqrt{n}$	rzeczywiste prawdopodobieństwo	rzeczywiste prawdopodobieństwo dlaklasy
0,1	−0,36995	0,10911	0,10911
0,2	−0,24296	0,20937	0,10026
0,3	−0,15138	0,30719	0,09782
0,4	−0,07314	0,40384	0,09665
0,5	0,0	0,50000	0,09616

Jeśli badaniu poddana zostanie seria ciągów pochodzących z idealnego generatora, to przeciętne częstości, otrzymywane dla poszczególnych klas, dążyć będą do przedstawionych powyżej prawdopodobieństw dla klas. Biorąc liczbę powtórzeń testu (liczność serii ciągów) wynoszącą 100, otrzymamy następujące średnie liczności dla poszczególnych klas:

10,911 10,026 9,782 9,665 9,616
9,616 9,665 9,782 10,026 10,911

Obliczone dla tych wartości prawostronne prawdopodobieństwo w teście zgodności równe jest 0,999999, co oznacza, że błąd określenia wariancji nie ma istotnego wpływu na otrzymywany wynik. Nawet jeśli zamiast wartości ułamkowych, przyjmiemy najbliższe liczby całkowite najlepiej pasujące do powyższych średnich:

$$11,\ 10,\ 10,\ 10,\ 9,\ 9,\ 10,\ 10,\ 10,\ 11,$$

to i tak różnica nie będzie znacząca.

Okazuje się jednak, że już dla liczby powtórzeń wynoszącej 430, liczone jak powyżej, prawdopodobieństwo spada poniżej 0,9, dla 610 – poniżej 0,5, zaś dla 980 – poniżej 0,01. Wynika z tego, że przeprowadzając test drugiego poziomu dla 1000 powtórzeń tego testu, niemal zawsze otrzymamy odpowiedź: "Twój generator jest bardzo, ale to bardzo, zły." Wyjątkiem mogą być sytuacje, gdy rozpatrywany generator jest naprawdę bardzo zły, gdyż wtedy może mieć zastosowanie ludowe porzekadło o dwóch minusach.

2. Testy autokorelacyjne dla ciągu binarnego

Zgodnie z tytułem, w rozdziale tym zajmiemy się badaniem ciągu binarnego $s^l = s_1, s_2, \ldots, s_l$ – ciąg binarny o długości l, gdzie $s_i \in \{0, 1\}$ i tworzą one ciąg niezależnych zmiennych losowych z prawdopodobieństwem 1/2 otrzymania 0.

2.1. Test dla ciągu bitów

Dla binarnego ciągu s^{n+t} obliczana jest empiryczna korelacja z jego niecyklicznym przesunięciem o t pozycji:

$$A_i = \sum_{i=1}^{n} (s_i \oplus s_{i+t}),$$

gdzie $\oplus$ oznacza alternatywę wykluczającą (sumę modulo 2).

Powyższa procedura jest tożsama z przeprowadzeniem testu momentów pierwszego rzędu dla bloku jednobitowego dla ciągu powstałego przez pobitowe dodanie modulo 2 ciągu wyjściowego i jego przesunięcia o t pozycji. Wobec powyższego statystyka testowa ma asymptotycznie rozkład określony formułą:

$$\Pr(A_t < x) = \Phi\left(\left(x - \frac{n}{2}\right)\frac{2}{\sqrt{n}}\right) - \frac{1}{12n}\varphi^{(3)}\left(\left(x - \frac{n}{2}\right)\frac{2}{\sqrt{x}}\right), \quad [3]$$

gdzie $\Phi(x)$ jest dystrybuantą rozkładu $N(0,1)$, a $\varphi^{(3)} = \frac{x(3-x^2)}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ jest trzecią pochodną gęstości tego rozkładu. Dla bardzo dużych n drugi składnik można pominąć, otrzymując w ten sposób przybliżenie rozkładem $N(n/2, \sqrt{n}/2)$.

2.2. Test dla ciągu bloków d bitowych

Testy opisane w punktach 1.2. i 2.1. są, z punktu widzenia obliczeń przy użyciu komputera cyfrowego, dwoma skrajnymi przypadkami – w drugim ciąg dzielony jest na bloki jednobitowe, a w pierwszym na, w teorii, nieskończonej długości, a w praktyce 32 bitowe. Tak długie bloki rodzą jednak dwa problemy:

- ciąg określonej długości można podzielić na stosunkowo niewiele takich bloków,
- najmniej znaczące bity, okazują się być praktycznie nieznaczące, ze względu na ich marginalny wpływ na wartość statystyki testowej.

Powyższe skłoniły nas do zaproponowania modyfikacji testu z punktu 1.2., która uwzględnia skończoną długość bloku.

Wyjściowy ciąg binarny $s^{d(n+t)}$ dzielony jest na rozłączne bloki długości d bitów, które następnie utożsamiamy z liczbami naturalnymi ze zbioru $\{0, 1, 2, \ldots, 2^d - 1\}$. W ten sposób otrzymujemy ciąg liczb całkowitych $C^{n+t} = c_1, c_2, \ldots, c_{n+t}$. Na podstawie założenia o równomierności i niezależności rozkładu bitów w ciągu wyjściowym, można również założyć, że tworzą one ciąg niezależnych zmiennych losowych o rozkładzie równomiernym na zbiorze $\{0, 1, 2, \ldots, 2^d - 1\}$.

Wyznaczana jest statystyka testowa:

$$S^d_{ncykl} = \sum_{i=1}^{n} c_i c_{i+t}, \quad t > 0.$$

Poniżej prezentujemy sposób wyznaczenia parametrów rozkładu i dystrybuanty statystyki testowej.

Zakładamy, że dla dużych n mieć ona będzie, w przybliżeniu, rozkład normalny. Aby zminimalizować błąd tego przybliżenia posłużymy się formułą: [3], [9]

$$\begin{aligned} F(x) = \Phi(x) &- \frac{1}{3!}\frac{\mu_3}{\sigma^3}\varphi^{(2)}(x) + \frac{1}{4!}\left(\frac{\mu_4}{\sigma^4} - 3\right)\varphi^{(3)}(x) \\ &- \frac{1}{5!}\left(\frac{\mu_5}{\sigma^5} - 10\frac{\mu_3}{\sigma^3}\right)\varphi^{(4)}(x) + \frac{10}{6!}\left(\frac{\mu_3}{\sigma^3}\right)^2\varphi^{(5)}(x) \\ &- \frac{35}{7!}\frac{\mu_3}{\sigma^3}\left(\frac{\mu_4}{\sigma^4} - 3\right)\varphi^{(6)}(x) - \frac{280}{9!}\left(\frac{\mu_3}{\sigma^3}\right)^3\varphi^{(8)}(x) + \ldots, \end{aligned}$$

gdzie: $F(x)$ – dystrybuanta znormalizowanej statystyki testowej, $\Phi(x)$ – dystrybuanta standardowego rozkładu normalnego, $\varphi^{(k)}$ – k-ta pochodna gęstości standardowego rozkładu normalnego, σ – odchylenie standardowe zmiennej S^d_{ncykl}, μ_k – moment centralny rzędu k zmiennej S^d_{ncykl}.

Oczywiście w prezentowanej powyżej formule jest nieskończenie wiele wyrazów, my jednak ograniczyliśmy się do wyznaczenia przedstawionych, co, jak pokażemy dalej, pomijając skrajne przypadki, jest znacznie nadmiarowe.

Istotnym ograniczeniem stosowalności powyższej formuły jest warunek niezależności składników sumy tworzącej zmienną losową – naszą statystykę testową. W oczywisty sposób warunek ten nie jest tutaj spełniony, w dalszej części pokażemy jednak, że wobec niewielkiej siły korelacji można ją zaniedbać.

Rozpoczynamy od wyznaczenie momentów rzędu od 1 do 4 zmiennej S^d_{ncykl}. Ponieważ rachunki temu towarzyszące są dość żmudne i cechują się znaczną objętością, pozwolimy sobie zaprezentować tok postępowania na przykładzie momentów rzędu 1 i 2, dla kolejnych podając jedynie wynik końcowy.

Dla momentu rzędu 1 mamy:

$$m_1 = E[S^d_{ncykl}] = E\left[\sum_{i=1}^{n} c_i c_{i+t}\right]^2 = \sum_{i=1}^{n} E[c_i c_{i+t}] = \sum_{i=1}^{n} E[c_i]E[c_{i+t}]$$
$$= n(E[c_i])^2 = \frac{(2^d-1)^2}{4} n.$$

Skorzystaliśmy z addytywności wartości oczekiwanej i niezależności zmiennych c_i i c_{i+t}.

Dla momentu rzędu 2 mamy:

$$m_2 = E\left[\sum_{i=1}^{n} c_i c_{i+t}\right]^2 = E\left[\sum_{i=1}^{n} (c_i c_{i+t})^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_i c_{i+t} c_j c_{j+t}\right].$$

Pierwszy składnik jest prosty i wyznaczamy go tak, jak dla momentu rzędu 1. Dla drugiego rozpatrzyć należy dwa przypadki: gdy $j = i + t$ oraz pozostałe. Tak więc dostajemy:

$$m_2 = nE[c_i^2 c_{i+t}^2] + 2(n-t)E[c_i c_{i+t}^2 c_{i+2t}] + (n^2 - 3n + 2t)E[c_i c_{i+t} c_j c_{j+t}]$$
$$= \frac{n}{36}(2^d-1)^2(2\cdot 2^d - 1)^2 + \frac{n-t}{12}(2^d-1)^3(2\cdot 2^d - 1)$$
$$+ \frac{n^2 - 3n + 2t}{16}(2^d-1)^4$$
$$= \frac{(2^d-1)^2}{144}(9(2^d-1)^2 n^2 + (13\cdot 2^{2d} + 2\cdot 2^d - 11)n$$
$$- 6(2^d-1)(2^d+1)t).$$

Przy pomocy dwóch powyższych momentów wyznaczamy wariancję:

$$\mu_2 = \sigma^2 = D^2[S^d_{ncykl}] = \frac{(2^d-1)^2(2^d+1)}{144}((13\cdot 2^d - 11)n - 6(2^d-1)t).$$

Dla kolejnych momentów i momentów centralnych mamy:

$$\begin{aligned} m_3 &= E\Bigg[\sum_{i=1}^{n}(c_ic_{i+t})^3 + 3\sum_{i=1}^{n}\sum_{j\neq i}(c_ic_{i+t})^2c_jc_{j+t} \\ &\quad + 6\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1}\sum_{k=j+1}^{n} c_ic_{i+t}c_jc_{j+t}c_kc_{k+t}\Bigg] \\ &= \frac{(2^d-1)^4}{192}\Big(3(2^d-1)^2n^3 + (2^d+1)(13\cdot 2^d-11)n^2 \\ &\quad - 2(2^d+1)(3(2^d-1)t - 4(2^d+1))n - 8(2^d+1)^2t\Big), \end{aligned}$$

$$\mu_3 = \frac{1}{24}(2^d-1)^4(2^d+1)^2(n-t),$$

$$\begin{aligned} m_4 = \frac{(2^d-1)^2}{2^83^35^2}\Big(&675(2^d-1)^6n^4 + 450(2^d-1)^4(2^d+1)(13\cdot 2^d-11)n^3 \\ &- 25(2^d-1)^2(2^d+1)(108(2^d-1)^3t - A_1)n^2 \\ &- 6(2^d+1)(50(2^d-1)^3(2^d+1)(37\cdot 2^d-35)t - A_2)n \\ &+ 40(2^d-1)(2^d+1)(45(2^d-1)^3(11\cdot 2^d-7)t - A_3)t\Big) \end{aligned}$$

$$\begin{aligned} A_1 &= (2^d+1)(457\cdot 2^{2d} - 862\cdot 2^d + 409), \\ A_2 &= 327\cdot 2^{5d} + 167\cdot 2^{4d} - 2606\cdot 2^{3d} + 2514\cdot 2^{2d} + 367\cdot 2^d - 753, \\ A_3 &= 215\cdot 2^{4d} + 130\cdot 2^{3d} - 792\cdot 2^{2d} + 110\cdot 2^d + 313, \end{aligned}$$

$$\begin{aligned} \mu_4 = \frac{(2^d-1)^2(2^d+1)}{2^83^35^2}\Big(&25(2^d-1)^2(2^d+1)(13\cdot 2^d-11)^2n^2 \\ &- 6(50(2^d-1)^3(2^d+1)(13\cdot 2^d-11)t - A_2)n \\ &+ 20(2^d-1)(45(2^d-1)^3(2^d+1)t - A_3)\Big)t \end{aligned}$$

$$
\begin{aligned}
m_5 = \frac{(2^d-1)^4}{2^{10}3^35}\Big(&175(2^d-1)^6n^5 + 150(2^d-1)^4(2^d+1)(13\cdot 2^d-11)n^4\\
&- 25(2^d-1)^2(2^d+1)(36(2^d-1)^3t - A_4)n^3\\
&- 2(2^d+1)(150(2^d-1)^3(2^d+1)(25\cdot 2^d-23)t - A_5)n^2\\
&+ 4(2^d+1)(225(2^d-1)^4(2^d+1)t^2 - 5(2^d-1)A_6t - 16(2^d+1)A_7)n\\
&+ 32(2^d+1)^2(75(2^d-1)^3(2^d+1)t + A_8)t\Big),
\end{aligned}
$$

$$
\begin{aligned}
A_4 &= (2^d+1)(313\cdot 2^{2d} - 574\cdot 2^d + 265),\\
A_5 &= 3581\cdot 2^{5d} - 1699\cdot 2^{4d} - 13018\cdot 2^{3d} + 70\cdot 2^d + 653,\\
A_6 &= 595\cdot 2^{4d} + 170\cdot 2^{3d} - 1512\cdot 2^{2d} + 70\cdot 2^d + 653,\\
A_7 &= 37\cdot 2^{4d} - 120\cdot 2^{3d} + 214\cdot 2^{2d} - 120\cdot 2^d - 23,\\
A_8 &= 59\cdot 2^{4d} - 330\cdot 2^{3d} + 668\cdot 2^{2d} - 330\cdot 2^d - 91,
\end{aligned}
$$

$$
\begin{aligned}
\mu_5 = \frac{(2^d-1)^4(2^d+1)^2}{2^63^35}\Big(&25(2^d-1)^2(2^d+1)(13\cdot 2^d-11)^2n^2\\
&- (25(2^d-1)^2(2^d+1)(19\cdot 2^d-17)t - A_9)n\\
&+ 2(75(2^d-1)^3(2^d+1)t + A_8)t\Big)
\end{aligned}
$$

$$
A_9 = 148\cdot 2^{4d} - 480\cdot 2^{3d} + 856\cdot 2^{2d} - 480\cdot 2^d - 92
$$

Należy zaznaczyć, że powyższe formuły będą prawdziwe dla $n \geq 4t$, co oznacza minimalną długość badanego ciągu wynoszącą $5t$. W dalszej części pokażemy, że dla dużych długości ciągu możliwe będzie dopuszczenie przesunięć równych połowie długości ciągu. Dopuszczenie większych t, w pierwszym przypadku, wymagałoby dodatkowych żmudnych rachunków nie wnosząc do testu nowej jakości, przekroczenie przez t połowy długości ciągu równałoby się pomijaniu części wyrazów.

Obliczone momenty pozwalają wyznaczyć wszystkie siedem, zaprezentowanych powyżej, składników rozwinięcia dystrybuanty. Posiłkując się nimi można również wyznaczyć współczynnik korelacji pomiędzy wyrazami sumy:

$$
corr(c_ic_{i+t}, c_jc_{j+t}) = \begin{cases} \dfrac{3(2^d-1)}{7\cdot 2^d-5}, & \text{dla } j = i \pm t,\\ 0, & \text{w p.p.} \end{cases}
$$

Przyjmuje on, jak widać, wartości uważane za przeciętne, należy jednak pamiętać, że dla określonego składnika niezerowa wartość pojawia się jedynie

w przypadku dwóch lub tylko jednego innego składnika, tak więc średnia korelacja jest odwrotnie proporcjonalna do długości ciągu.

Jeśli w powyższych rozważaniach sumy zastąpimy odpowiednimi całkami, to wyznaczymy parametry dla testu z punktu 1.2. To samo można uzyskać dzieląc powyższe wyrażenia przez odpowiednią potęgę $(2^d - 1)$ i przechodząc z d do nieskończoności.

Drugim istotnym spostrzeżeniem jest, że znany z literatury test zaprezentowany w punkcie 2.1. nie jest szczególnym przypadkiem powyższego. Stał by się nim, gdyby sumę modulo 2 zastąpić iloczynem.

W kolejnym rozdziale zajmiemy się ustaleniem minimalnych wartości parametrów n i d, dla których uzasadnione jest posługiwanie się podanym przybliżeniem, jak również zastanowimy się nad koniecznością wykorzystywania w nim wszystkich siedmiu składników.

Na koniec zbadamy relację pomiędzy dwoma przedstawionymi powyżej testami.

3. Analiza własności statystyki testu dla ciągu bloków d bitowych

3.1. Istotność składników rozwinięcia

W punkcie tym sprawdzimy, czy wyznaczone w poprzednim punkcie przybliżenie dystrybuanty jest wystarczające.

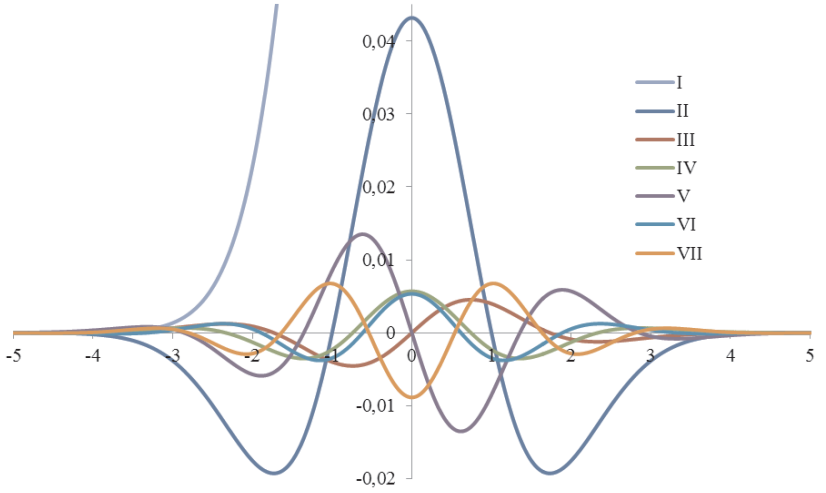
Jako jedną z miar wpływu każdego ze składników na wartość formuły przyjęliśmy największą, co do modułu, wartość przez niego przyjmowaną. Ponieważ parametry n, t i d mają wpływ jedynie na wartość współczynnika stojącego przed właściwą funkcją, więc niezależnie od nich można analitycznie wyznaczyć tę maksymalną wartość. W tabeli 2 zebraliśmy tak otrzymane wartości dla różnych kombinacji parametrów. Jeśliby ograniczyć się jedynie do tego wskaźnika, to bez wahania można by ograniczyć się do dwóch pierwszych składników, gdyż jedynie dla skrajnie małych n „poprawka" dawana przez największy z pozostałych przekracza kilka procent wartości wnoszonej przez drugi. Okazuje się, że równie istotne jak sama wartość maksymalna, jest również położenie argumentu jej odpowiadającego – w czterech przypadkach jest nim 0, w dwóch pozostałych są to wartości około $\pm 0{,}617$ i $\pm 0{,}742$, a więc wszystkie one położone są w centralnej części wykresu dystrybuanty, gdzie nad wszystkimi dominuje pierwszy składnik.

Wykresy wszystkich składników, odpowiadające pierwszej z umieszczonych w tabeli 2 kombinacji parametrów, zamieściliśmy na rysunku 1.

Tabela 2

Maksymalne wartości bezwzględne uzyskiwane przez poszczególne składniki dla różnych kombinacji parametrów n, t i d

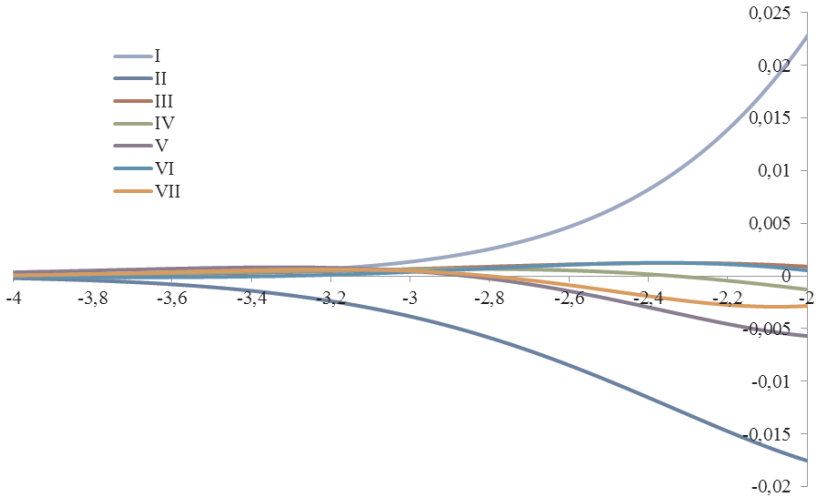
$n/t/d$	10/1/1	1000/1/1	1000/1/32	1000/100/32	$10^6/1/1$	$10^6/1/32$
max|II|	0,0432	0,0045117	0,0032288	0,0031203	$1{,}4 \cdot 10^{-4}$	$1{,}0 \cdot 10^{-4}$
max|III|	0,0045	0,0000568	0,0000319	0,0000274	$5{,}7 \cdot 10^{-8}$	$3{,}2 \cdot 10^{-8}$
max|IV|	0,0057	0,0000054	0,0000022	0,0000023	$1{,}7 \cdot 10^{-10}$	$7{,}0 \cdot 10^{-11}$
max|V|	0,0135	0,0001475	0,0000756	0,0000706	$1{,}5 \cdot 10^{-7}$	$7{,}6 \cdot 10^{-8}$
max|VI|	0,0053	0,0000070	0,0000028	0,0000023	$2{,}2 \cdot 10^{-10}$	$8{,}9 \cdot 10^{-11}$
max|VII|	0,0089	0,0000101	0,0000037	0,0000033	$3{,}2 \cdot 10^{-10}$	$1{,}2 \cdot 10^{-10}$
$\frac{\max\|V\|}{\max\|II\|}$	0,3130	0,0327	0,0234	0,0226	0,0010	0,0007



Rysunek 1. Wykresy wszystkich wyznaczonych składników dla kombinacji parametrów 10/1/1

W przypadku pierwszego z nich jest to jedynie mały fragment, gdyż dostosowaliśmy zakres osi rzędnych do pozostałych, dystrybuanta standardowego rozkładu normalnego jest zaś powszechnie znana. Na rysunku tym, oprócz licznych maksimów i minimów lokalnych poszczególnych funkcji i ich wzajemnej relacji, zwrócić należy uwagę na jego lewy kraniec, który osobno przedstawiliśmy na rysunku 2.

W wielu przypadkach, jak chociażby podczas pojedynczego badania, kiedy weryfikujemy hipotezę odnośnie jednego zadanego ciągu, istotniejsze od maksymalnych odchyleń w centralnej części stają się tzw. „ogony" rozkładu. Właśnie w tej części kolejne składniki mogą przeważać nad pierw-

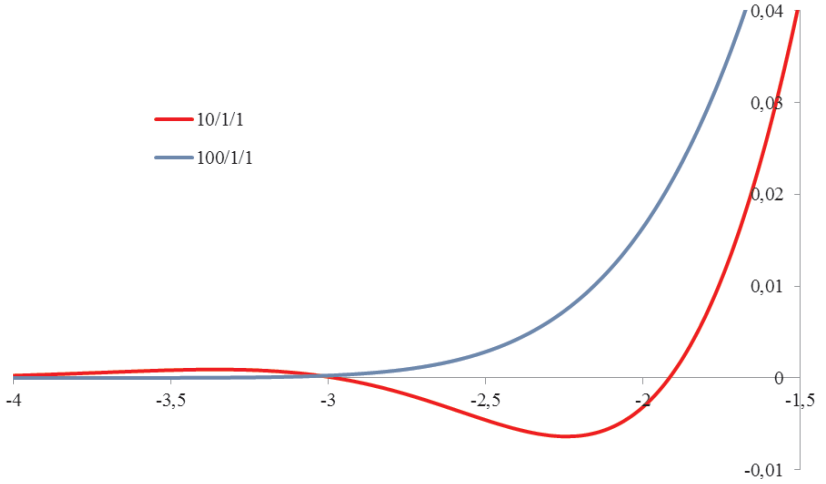


Rysunek 2. Fragmenty wykresów wyznaczonych składników dla kombinacji parametrów 10/1/1

szym i takie zjawisko obserwujemy na rysunku 2. W tym konkretnym przypadku okazuje się, że otrzymana formuła jest nieprawidłowa, a ściślej rzecz biorąc, że dla tak małej długości ciągu, należałoby użyć większej liczby wyrazów rozwinięcia. Prowadzi to do otrzymania funkcji niemonotonicznej, a dodatkowo przyjmującej wartości ujemne, a taka nie może być dystrybuantą. Dobrze widać to na rysunku 3, gdzie zamieściliśmy wykres funkcji będącej sumą wszystkich siedmiu wyrazów rozwinięcia. Na tym samym rysunku nanieśliśmy również wykres funkcji odpowiadającej kombinacji 100/1/1, dla której dopiero analiza danych liczbowych pozwala dostrzec opisane wyżej nieprawidłowości, są one jednak na tyle małe, że nie wpływają na jakość otrzymywanych wyników.

Analogiczne rozważania można powtórzyć dla prawej części rozkładu, z tą tylko różnicą, że rozpatrywać należy odległość otrzymywanych wartości od 1, nie od 0.

Na drugim biegunie wykorzystania testu znajduje się badanie generatorów, w którym zakładamy, że dla źródła losowego otrzymywać będziemy ciąg prawdopodobieństw wybieranych, poprzez statystykę testową, z rozkładem równomiernym na przedziale $(0, 1)$. Dla takiego ciągu powszechnie stosowany jest następnie klasyczny test Kołmogorowa – Smirnowa [6] lub jeden z testów z rodziny testów Andersona – Darlinga [1].



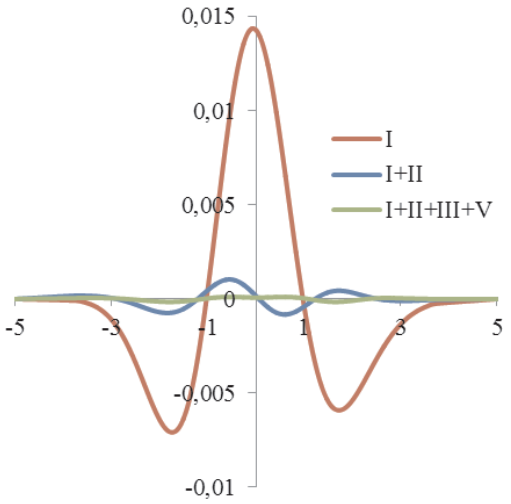
Rysunek 3. Fragment wykresu rozwinięcia do siódmego wyrazu dla kombinacji parametrów 10/1/1 i 100/1/1

Dla naszych celów najwygodniejsze będzie odwołanie się do testu Kołmogorowa – Smirnowa, w którym miarą jakości jest maksymalne odchylenie dystrybuanty empirycznej od teoretycznej. Przedstawiona poniżej procedura, nie jest zgodna z założeniami testu, więc nie uprawnia do uzyskania ścisłych wskaźników liczbowych, tym niemniej pozwoli wychwycić interesujące nas zależności.

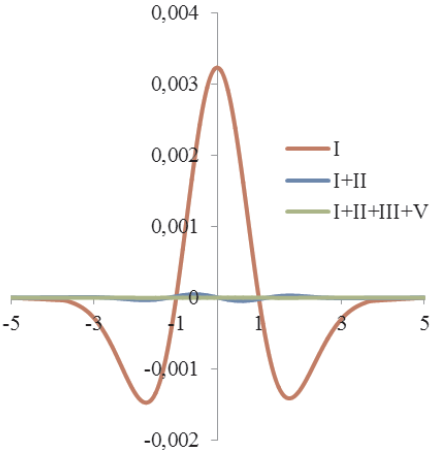
Wykorzystaliśmy trzy warianty dystrybuanty, za podstawę przyjęliśmy rozwinięcie składające się z siedmiu wyrazów oraz wersje zredukowane tylko do pierwszego (I), dwóch pierwszych wyrazów (I + II), czterech wyrazów (I + II + III + V). Naszym celem była ocena możliwości wykorzystania tych zredukowanych przybliżeń.

Dla każdego z trzech powyższych wariantów wyznaczaliśmy kolejne kwantyle, z krokiem 10^{-5}, a następnie dla nich obliczaliśmy prawdopodobieństwo według pełnej formuły. Na rysunkach od 4 do 6 prezentujemy odchylenia pomiędzy otrzymanymi prawdopodobieństwami a wartościami wyjściowymi, dla których zostały otrzymane, dla n wynoszących 100, 1000 i 10 000 oraz t i d równych 1. Na osi poziomej odkładane są kwantyle odpowiadające przypadkowi (I).

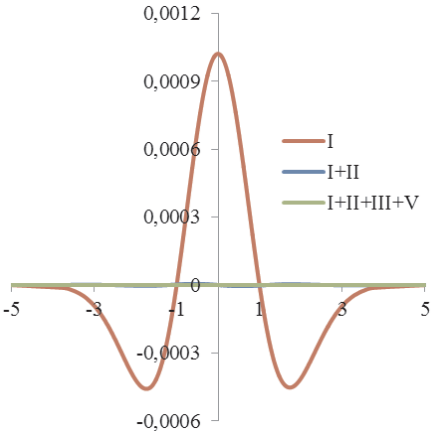
Wykresy są bardzo podobne, szczególnie w przypadku przebiegu (I), gdzie, pomijając symetryzowanie się, różnią się jedynie skalą. To samo dotyczy przebiegów dla pozostałych dwóch wariantów. Zauważyć należy, że gasną one wyraźnie szybciej od pierwszego.



Rysunek 4. Wykres odchyleń dla $n = 100$

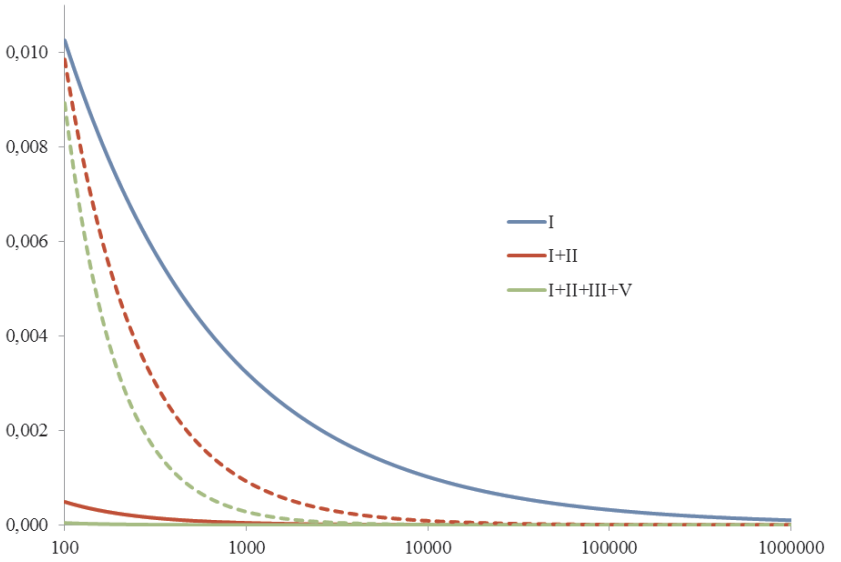


Rysunek 5. Wykres odchyleń dla $n = 1000$



Rysunek 6. Wykres odchyleń dla $n = 10\,000$

Na rysunku 7 prezentujemy zależność pomiędzy długością ciągu, a maksymalnym odchyleniem dla poszczególnych wariantów. Dodatkowo, aby uwypuklić różne tempa gaśnięcia tych odchyleń, dla poszczególnych wariantów, liniami przerywanymi oznaczyliśmy przebieg dla wariantu $(\mathrm{I}+\mathrm{II})$ po przemnożeniu przez 20, zaś dla wariantu $(\mathrm{I}+\mathrm{II}+\mathrm{III}+\mathrm{V})$ po przemnożeniu przez 200. Wykres ten pokazuje, że wariant wykorzystujący jedynie pierwszy wyraz osiąga dokładność uzyskiwaną przez dwa pozostałe dla



Rysunek 7. Maksymalne odchylenia notowane dla poszczególnych wariantów w zależności od długości ciągu

$n = 100$, dopiero dla ciągów o długości ok. 44 000 dla drugiego wariantu oraz przeszło 5 000 000 bloków dla trzeciego. Z kolei drugie przybliżenie daje błąd większy od 10 do przeszło 100 razy od trzeciego, przy czym górna wartość nie jest granicą, a wynika jedynie z zakresu przeprowadzonych obliczeń, co jednak ważniejsze ta ogromna różnica względna przekłada się na pomijalne: $4{,}5 \cdot 10^{-8}$ różnicy bezwzględnej. Wydaje się, że nawet w przypadkach wymagających bardzo dokładnych wyników, już dla ciągów złożonych z 1000 bloków można ograniczyć się do formuły wykorzystującej jedynie dwa pierwsze wyrazy.

Podsumowując ten punkt musimy przyznać, że dla bardzo małych n – rzędu dziesiątek, jeśli potrzebujemy precyzyjnych wyników dla ogonów rozkładu, podane przez nas przybliżenie okazuje się niewystarczające. Dla ciągów złożonych z setek wyrazów musimy zaś korzystać z proponowanego siedmiowyrazowego rozwinięcia, które dalej można ograniczyć do pięciu wyrazów. Jeśli weźmiemy pod uwagę maksymalne różnice pomiędzy wartościami uzyskiwanymi przy użyciu poszczególnych wariantów, to okazuje się, że nawet dla ciągów do 1000 bloków, wystarczające jest rozwinięcie wykorzystujące 4 wyrazy, zaś dla dłuższych, można ograniczyć się do dwóch pierwszych wyrazów.

3.2. Zbieżność składników rozwinięcia

Powyżej nie zaprezentowaliśmy formuły określającej dystrybuantę, a jedynie podaliśmy wystarczające do tego momenty centralne. Nigdy też nie wyznaczyliśmy jej explicite, zawsze posługiwaliśmy się rozbiciem na poszczególne wyrazy, gdyż w całości byłaby ona zbyt chaotyczna i nie wnosiłaby nic do rozważań. Tu również rozważać będziemy każdy z wyrazów z osobna.

Naszym pierwszym krokiem będzie spojrzenie na to w jaki sposób zależą współczynniki kolejnych wyrazów od n. W części przypadków posłużyliśmy się uproszczeniem polegającym na prezentowaniu jedynie dominującego składnika. Wybrane wyniki prezentujemy w tabeli 3. Część podawanych wartości liczbowych jest jedynie dziesiętnymi przybliżeniami.

Tabela 3
Dominujący składnik opisujący wpływ parametru n na współczynniki kolejnych wyrazów

t/d	1/1	1/32	$n/4/1$	$n/4/32$
wyraz II	$-\frac{4(n-1)}{(5n-2)^{3/2}}$	$-0{,}256/\sqrt{n}$	$-\frac{2\sqrt{2}}{9\sqrt{n}}$	$-0{,}231/\sqrt{n}$
wyraz III	$\frac{9(31n-82)}{(5m-2)^2}$	$0{,}058/n$	$\frac{7}{162n}$	$0{,}0335/n$
wyraz IV	$-\frac{4(2n-1)}{(5n-2)^{5/2}}$	$0{,}0583/n\sqrt{n}$	$\frac{28\sqrt{2}}{243n\sqrt{n}}$	$0{,}0643/n\sqrt{n}$
wyraz V	$\frac{8(n-1)^2}{(5n-2)^3}$	$0{,}0328/n$	$\frac{4}{81n}$	$0{,}0266/n$
wyraz VI	$-\frac{(n-1)(31n-82)}{3(5n-2)^{7/2}}$	$-0{,}0149/n\sqrt{n}$	$-\frac{7\sqrt{2}}{729n\sqrt{n}}$	$-0{,}00774/n\sqrt{n}$
wyraz VII	$-\frac{32(n-1)^3}{3(5n-2)^{9/2}}$	$-0{,}0028/n\sqrt{n}$	$-\frac{8\sqrt{2}}{2187n\sqrt{n}}$	$-0{,}00205/n\sqrt{n}$

Dla każdego z wyrazów analizujemy cztery kombinacje wartości parametrów t i d, przyjmujących swoje skrajne wartości. W przypadku t są to 1 i $n/4$, dla d przyjęliśmy 1 i 32. Ostatnia jest wartością czysto arbitralną, ale też oczywistą.

We wszystkich przypadkach współczynniki zbiegają do 0 wraz ze wzrostem n. Tempo zbiegania jest najsłabsze dla drugiego wyrazu, co doskonale zgadza się z wynikami zaprezentowanymi w tabeli 2.

Wyniki zawarte w tabeli 3 były podstawą do wyboru wariantów, które rozpatrywaliśmy w poprzednim punkcie – decydował wykładnik potęgi n dominującego składnika. Przebiegi na rysunku 7 niemal idealnie odpowiadają funkcjom proporcjonalnym do $1/\sqrt{n}$, $1/n$ oraz $1/n\sqrt{n}$, a więc dominującym składnikom odrzuconych wyrazów.

Powyższe potwierdza, że poszczególne składniki, wraz ze wzrostem długości ciągu, kolejno przestają istotnie wpływać na wyznaczaną wartość, zaś dla bardzo dużych n można ograniczyć się jedynie do pierwszego składnika. Dodatkowym zyskiem, w takim przypadku, jest możliwość rozszerzenia zakresu wartości przesunięcia t do $n/2$.

3.3. Zgodność

W punkcie 2.2. zaznaczyliśmy, że stosowane przybliżenie rozkładem normalnym, nawet wykorzystujące dodatkowe składniki, będzie prawdziwe dla dużych n. Wynika to z faktu, że elementy sumy nie mają rozkładu normalnego, nawet w przybliżeniu. Konsekwencje tego pokazujemy na poniższym przykładzie.

Załóżmy, że chcielibyśmy przeprowadzać test dla ciągów od długości 31 bitów. Możemy przyjąć następujące wartości parametrów $n = 30$, $t = 1$, $d = 1$, wówczas, według wyprowadzenia z punktu 2.2., dystrybuanta ma postać:

$$P\left(S^d_{ncykl} < \left(\frac{15}{2} + \frac{\sqrt{37}}{2}x\right)\right) \approx \Phi(x) + \big(2613{,}9 - 855{,}8x - 2408{,}6x^2 + 699{,}3x^3 - 213{,}7x^4 - 82{,}8x^5 + 41{,}5x^6 - 1{,}8x^8\big)\frac{e^{\frac{x^2}{2}}}{10^5}$$

przy czym wszystkie wartości zostały zaokrąglone do jednego miejsca po przecinku.

W rzeczywistości statystka testowa przyjmuje zaledwie 31 wartości całkowitych z zakresu $0 \ldots 30$. Tabela 4 zawiera funkcję prawdopodobieństwa dla tej dyskretnej zmiennej losowej, w drugiej kolumnie umieściliśmy prawdopodobieństwa uzyskiwane przy użyciu powyższej formuły. Z kolei na rysunku 8 zamieszczamy wykres dystrybuanty rzeczywistej i dla porównania wykres pokazanej powyżej funkcji z zaznaczeniem wartości odpowiadających argumentom całkowitoliczbowym. Zarówno dane w tabeli, jak i wykresy na rysunku pokazują, że powyższa formuła jest złym przybliżeniem rzeczywistego rozkładu. Jest praktycznie niemożliwe, aby dobry generator przeszedł test wykorzystujący przedstawione „przybliżenie", gdy testem drugiego poziomu będzie test Kołmogorowa-Smirnowa. Jeśli jednak w tej roli wykorzystamy test zgodności Pearsona, to dopiero dla kilkuset powtórzeń zaobserwujemy pogarszanie się wyników.

Okazuje się, że na dokładność przybliżenia wpływa nie tylko liczba bloków n, ale też liczba bitów składających się na pojedynczy blok d.

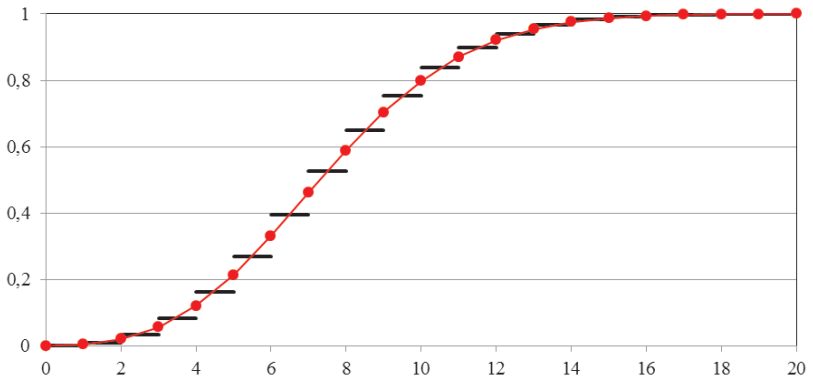
Tabela 4

Funkcja prawdopodobieństwa dla przypadku $n = 30$, $t = 1$, $d = 1$ i prawdopodobieństwa uzyskane na podstawie przybliżenia rozkładem normalnym

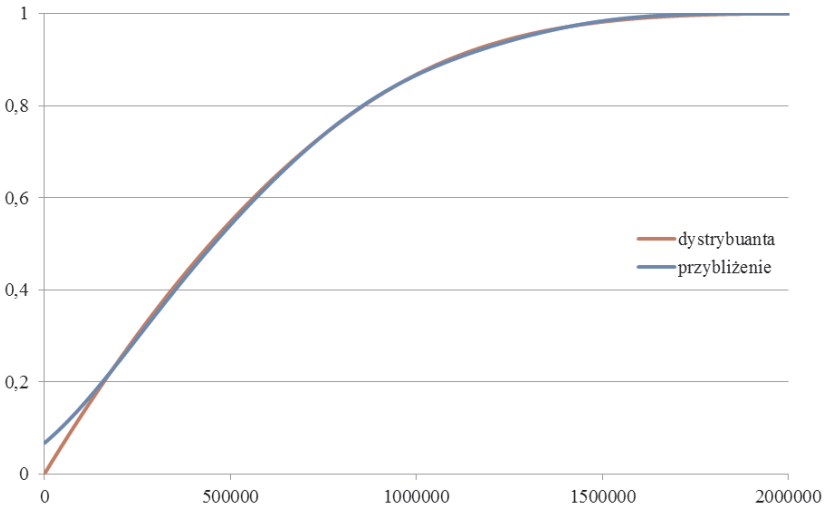
$S^d_{ncykl} = i$	p_i	$P(i-1 \leq S^d_{ncykl} < i)$	$\frac{\lvert p_i - P(i-1 \leq S^d_{ncykl} < i)\rvert}{p_i} \cdot 100\%$
0	0,00164	0,00021	87,5
1	0,00857	0,00491	42,7
2	0,02432	0,01596	34,4
3	0,04928	0,03604	26,9
4	0,07927	0,06381	19,5
5	0,10706	0,09352	12,6
6	0,12557	0,11737	6,5
7	0,13080	0,12941	1,1
8	0,12292	0,12787	4,0
9	0,10541	0,11482	8,9
10	0,08321	0,09452	13,6
11	0,06085	0,07179	18,0
12	0,04143	0,05067	22,3
13	0,02636	0,03349	27,0
14	0,01572	0,02082	32,4
15	0,00881	0,01214	37,8
16	0,00464	0,00658	41,9
17	0,00230	0,00331	43,6
18	0,00108	0,00155	44,0
19	0,00047	0,00070	47,4
≥ 20	0,00031	0,00053	70,9

Rozważmy przypadek $n = 2$, $t = 1$, $d = 10$, a więc odpowiadający ciągowi niemal tej samej długości, jak powyżej. Jak widać na rysunku 9, w znacznej części wykres dystrybuanty pokrywa się niemal dokładnie z wykresem funkcji wynikającej z przybliżenia. Jedynie dla początkowego fragmentu, odpowiadającego około 20% przypadków, różnica jest znacząca, podczas gdy dla takiego n moglibyśmy spodziewać się wszystkiego poza zgodnością.

Jeśli zwiększymy długość ciągu do, w dalszym ciągu bardzo małej wartości, 310 bitów i przyjmiemy $n = 30$, $t = 1$, $d = 10$, to okaże się, że dopiero dla ponad 180 powtórzeń testu, w teście Kołmogorowa-Smirnowa otrzymamy systematyczny błąd na poziomie 5%, dla 410 powtórzeń – 10%. Wyniki te otrzymaliśmy na podstawie przeprowadzenia 10^9 powtórzeń testu dla ciągów pochodzących z rejestru LFSR maksymalnego okresu o długo-



Rysunek 8. Dystrybuanta dla przypadku $n = 30$, $t = 1$, $d = 1$ oraz funkcja uzyskana na podstawie przybliżenia rozkładem normalnym

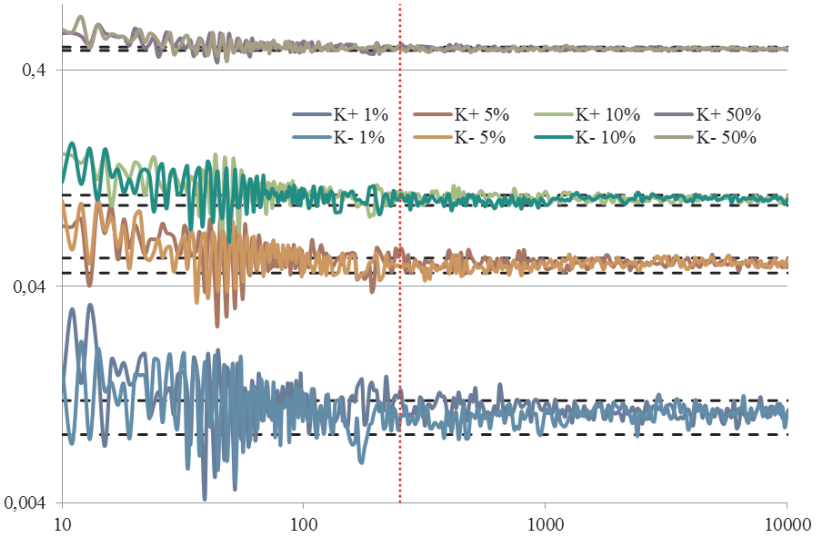


Rysunek 9. Dystrybuanta dla przypadku $n = 2$, $t = 1$, $d = 10$ oraz funkcja uzyskana na podstawie przybliżenia rozkładem normalnym

ści 128 bitów oraz algorytmu TRIVIUM. Zastosowaliśmy te dwa źródła ze względu na oferowaną przez nie szybkość generacji, jak również dobre własności statystyczne – udowodnione w przypadku LFSRa [4] i powszechnie przebadane w drugim przypadku. Źródła te wykorzystywane były również w dalszej części, we wszystkich przypadkach dawały one zgodne wyniki, innymi słowy przy użyciu omówionych tu testów atak odróżniający zakończył się niepowodzeniem, co potwierdza zasadność dokonanego doboru.

Ostatecznie postawiliśmy sobie pytanie, jaka powinna być długość badanego ciągu, aby otrzymywane wyniki uznać za wiarygodne. Wybraliśmy dwa warianty użycia testu, w pierwszym wykonywana jest niewielka liczba powtórzeń – u nas 10, w drugim bardzo duża – 10 000, następnie rozkład uzyskanych prawdopodobieństw przyrównywany jest do rozkładu równomiernego na przedziale $[0, 1]$ przy użyciu statystyk K^+ i K^- Kołmogorowa – Smirnowa, pełniących rolę testu drugiego poziomu. Dla danej kombinacji d i n powyższa procedura powtarzana była, dla każdego ze źródeł, 10 000 razy, w celu oszacowania częstości, z jaką w teście drugiego poziomu otrzymywane są prawdopodobieństwa nie większe od: 1%, 5%, 10% i 50%.

Na rysunkach 10 i 11 przedstawiliśmy wykresy częstości uzyskanych dla poszczególnych progów, dla jednobitowego bloku, w zależności od dłu-



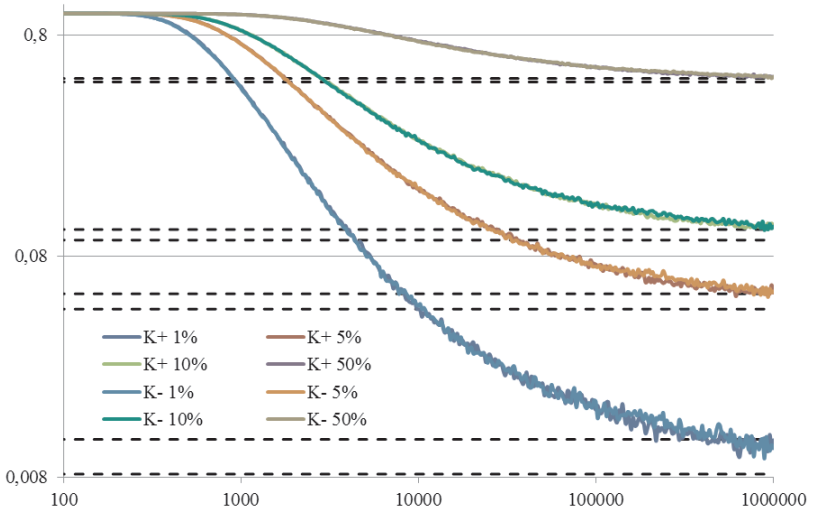
Rysunek 10. Częstości, z jakimi w teście drugiego poziomu otrzymywano prawdopodobieństwa nie większe od 1, 5, 10 i 50% dla $t = 1$ i $d = 1$, w zależności od długości ciągu, przy liczbie powtórzeń 10

gości badanego ciągu. Rysunek 10 przedstawia obraz dla liczby powtórzeń 10, zaś rysunek 11 dla 10 000. Linie przerywane ograniczają pasy odpowiadające 99% obszarowi centralnemu w schemacie Bernoulliego, dla każdego z przypadków. Obie osie mają skalę logarytmiczną.

Jak widać, począwszy od $n = 250$ (pionowa czerwona linia), możemy już mówić o niewielkich odchyleniach, zaś od 1000 o dobrej zgodności, a tym samym jakości, uzyskiwanych wyników.

Po znormalizowaniu przebiegi dla wszystkich czterech przypadków wykazują bardzo dużą zgodność.

Inny obraz uzyskujemy, gdy liczba powtórzeń testu wynosi 10 000. Otrzymane przez nas przebiegi prezentujemy na rysunku 11.



Rysunek 11. Częstości, z jakimi w teście drugiego poziomu otrzymywano prawdopodobieństwa nie większe od 1, 5, 10 i 50% dla $t = 1$ i $d = 1$, w zależności od długości ciągu, przy liczbie powtórzeń 10000

Na podstawie powyższego wykresu przyznać należy, że dla jednobitowego bloku przedstawione przez nas przybliżenie dystrybuanty zaczyna być akceptowalne dopiero dla ciągów o długości miliona bitów. Dla krótkich ciągów, przy dużej liczbie powtórzeń, test drugiego poziomu wskaże na istotne odchylenie od losowości – np. dla ciągu o długości 1000 bitów mamy ok. 50% prawdopodobieństwo otrzymania wyniku poniżej 0,01, nawet, jeśli badany jest „dobry" generator.

Dla większych wartości d otrzymywane przebiegi są bardzo podobne, przy czym zmienia się skala osi odciętych. W tabeli 5 przedstawiliśmy minimalne długości ciągu, dla których uznaliśmy, że test Kołmogorowa-Smirnowa nie wykrywa błędu wyznaczenia rozkładu statystyki testowej.

W trakcie obliczeń przyjęliśmy, że minimalną wartością n jest 5, stąd też wynikają wartości otrzymane dla 10 powtórzeń. Dla $d \geq 5$ wystarczające są długości ciągów pozwalające uzyskać kilka bloków.

Tabela 5
Minimalne długości ciągu wymagane do uzyskiwania poprawnych wyników w teście drugiego poziomu

d	10 powtórzeń testu	10 000 powtórzeń testu
1	250	1 000 000
2	20	150 000
5	30	100
10	60	200
20	120	440
30	180	570

W przypadku 10 000 powtórzeń widać, że bloki 1 i 2 bitowe sprawiają poważne problemy, które wynikają z dyskretnego charakteru obliczanej statystyki – w przeciwieństwie do przyjętego przez nas ciągłego przybliżenia. Test Kołmogorowa-Smirnowa wykrywa dla małych n całkowicie prawidłowe „schodki", takie jak te, które widać na rysunku 8. Dopiero na dużych n lub d stają się one na tyle drobne i gęste, by wystarczająco mogły być przybliżone linią ciągłą. Jak duże znaczenie ma w tym przypadku d można sobie uzmysłowić porównując liczbę wartości przyjmowanych przez statystykę testową dla różnych kombinacji d i n. Na przykład dla $d = 5$ i $n = 19$, co oznacza ciąg o długości 100 bitów, statystyka testowa przyjmuje wartości z zakresu od 0 do 18 259 (przy czym nie wszystkie wartości pośrednie są możliwe). Dla $d = 1$ uzyskanie takiej rozpiętości wymaga ciągu prawie 183 razy dłuższego. Z kolei dla $d = 30$ już dwa bloki dadzą zakres $6 \cdot 10^{13}$ razy większy, w tym jednak przypadku $n = 1$ jest stanowczo zbyt małe, by zaczęło działać centralne twierdzenie graniczne.

Podsumowując ośmielamy się stwierdzić, że zaproponowane przybliżenie jest satysfakcjonująco dobre nawet dla małych d i bardzo dobre dla większych. Co więcej, wymagana dla przypadku $d = 1$, przy bardzo dużej liczbie powtórzeń, długość ciąguwynosząca milion bitów nie jest przesadnie duża – jest ona używana chociażby w pakiecie Statistical Test Suite [8], zaś w pakiecie DIEHARD [7] jest wielokrotnie większa.

4. Czy potrzebujemy tego testu

W rozdziale tym pokażemy dwa powody, które sprawiają, że proponowany test można traktować jako coś więcej aniżeli tylko kolejną wariację na temat autokorelacji ciągu.

4.1. Jest test z niecyklicznym przesunięciem

Zgodnie z tytułem tego punktu, nie ukrywamy, że test dla ciągu liczbowego, przedstawiony w punkcie 1.2., był punktem wyjścia do naszych rozważań, a bezpośrednim impulsem do ich przeprowadzenia było spostrzeżenie przedstawione poniżej.

Rozważmy przykład, w którym przeprowadzamy test dla ciągu o długości 3000 bitów, który dzielony jest na bloki 30 bitowe, przesunięcie wynosi 1 blok.

Statystyka testu z punktu 1.2. ma wówczas rozkład

$$N\left(0, \sqrt{\frac{13}{144 \cdot 99} - \frac{1}{24 \cdot 99^2}}\right),$$

zaś w zaproponowanym przez nas teście dostajemy:

$$P(S^d_{ncykl} < (E+\sigma x)) \approx \Phi(x) + \big(10275 - 1275{,}9x - 10170x^2 + 1081{,}5x^3 - 123{,}1x^4 - 131{,}23x^5 + 25{,}495x^6 - 1{,}122x^8\big)\frac{e^{-\frac{x^2}{2}}}{10^6},$$

gdzie

$$E = \frac{(2^{30}-1)^2 \cdot 99}{4} \quad \text{i} \quad \sigma = \frac{(2^{30}+1)(2^{30}-1)^2(1281 \cdot 2^{30} - 1083)}{144}.$$

W obu przypadkach dostajemy niemal równoważne formuły, różniące się jedynie brakiem standaryzacji wartości przypisywanych blokom oraz niewielką poprawką wnoszoną przez kolejne składniki rozwinięcia.

Poszukiwać będziemy takich minimalnych wartości pojedynczego bloku, które prowadzą do przyjęcia przez statystykę testową wartości równych kwantylom rzędów: 0,99, 0,5, 0,01 oraz 10^{-6}. Kwantyle te to: $1{,}9805 \cdot 2^{64}$, $1{,}5469 \cdot 2^{64}$, $1{,}1132 \cdot 2^{64}$ oraz $1{,}3216 \cdot 2^{63}$. Zakładając, że wszystkie bloki mają tę samą wartość, to jest nią odpowiednio:

$$1{,}1315 \cdot 2^{29}, \quad 2^{29}, \quad 1{,}6966 \cdot 2^{28} \quad \text{oraz} \quad 1{,}3072 \cdot 2^{28}.$$

Wszystkie one są większe od 2^{28}, co oznacza, że ich reprezentacja binarna jest liczbą co najmniej 29 bitową. Oznacza to też, że jeśli w każdym bloku dwa najbardziej znaczące bity będą równe 0, to wartość pozostałych nie będzie miała praktycznie żadnego znaczenia – nie uzyskamy prawdopodobieństwa większego do 10^{-6}. Te dwa bity, na każde 30 mają większe znaczenie, aniżeli pozostałych 28.

Aby uzyskać ścisłą miarę istotności poszczególnych bitów, wyznaczyliśmy wartość oczekiwaną wartości bezwzględnej zmiany prawdopodobieństwa uzyskiwanego w teście, na skutek zmiany wartości bitu znajdującego się na wybranej pozycji. Użycie wartości bezwzględnej wynika z faktu, że obserwowane zmiany mają, dla małych różnic, symetryczny charakter, a więc brane wprost kompensowałyby się do zera.

Do wyznaczania interesującej nas miary posłużyliśmy się formułą:

$$\nabla_1 = \int_0^1 \frac{f(F^{-1}(p))}{2^{2d+1}} \sum_{a=0}^{2^d-1} \sum_{b=0}^{2^d-1} (F(F^{-1}(p)+(a+b)2^i) - F(F^{-1}(p)-(a+b)2^i))dp,$$

gdzie $F(\cdot)$ jest dystrybuantą rozkładu statystyki testowej, a $f(\cdot)$ jego gęstością, zaś $i = 0, 1, \ldots, d-1$ jest pozycją bitu, którego istotność rozpatrujemy. W przedstawionych tu rozważaniach ograniczamy się jedynie do bitów z „wewnętrznych" bloków, tzn. takich, które pojawiają się dwukrotnie w sumie będącej statystyką testową. W przypadku bitów z „zewnętrznych" bloków wystarczy wszystkie prezentowane wyniki podzielić przez 2.

Powyższe całkowanie wykonaliśmy numerycznie, posiłkując się kwantylami rozkładu normalnego, a dla $d > 15$ wartości sumy $(a + b)$ przeglądaliśmy z krokiem 256 (przeprowadziwszy uprzednio sprawdzenie, że nie powoduje to istotnej zmiany uzyskiwanych wyników). W tabeli 6 zebraliśmy wyniki uzyskane dla wybranych długości bloku, założyliśmy stałą długość ciągu, wynoszącą 600 bitów i przesunięcie równe 1 blok.

Dane zawarte w tabeli 6 pokazują, jak ogromne są dysproporcje wpływu bitów znajdujących się na poszczególnych pozycjach w bloku. Dla ustalonej długości bloku d, wpływ każdego kolejnego, mniej znaczącego, bitu jest dwukrotnie mniejszy od poprzedniego. Iloraz ten jest stały i niezależny od d. Prowadzi to do prostej konkluzji, że najbardziej znaczące bity poszczególnych bloków wpływają na wartość otrzymywanego w teście prawdopodobieństwa w równym (a ściśle biorąc nieco większym) stopniu, co wszystkie pozostałe razem wzięte.

Rozważania z początku tego punktu możemy uzupełnić o stwierdzenie, że nie tylko w skrajnych przypadkach dwa najbardziej znaczące bity, w 30 bitowych blokach, decydują o wartości uzyskanego prawdopodobieństwa, ale też średnio biorąc ich wpływ przewyższa trzykrotnie wpływ wszystkich pozostałych.

Drugą obserwacją jest narastający wpływ najbardziej znaczącego bitu, towarzyszący wzrostowi długości bloku. Na rysunku 12 przedstawiliśmy wykresy zależności przeciętnego wpływu najbardziej znaczącego bitu bloku na wartość prawdopodobieństwa w funkcji długości bloku.

Tabela 6
Średni wpływ poszczególnych bitów w bloku na wartość prawdopodobieństwa w teście

$d = 2$	
i	$n = 299$
0	0,0188056
1	0,0375535

$d = 5$	
i	$n = 119$
0	0,0052052
1	0,0104101
2	0,0208171
3	0,0416104
4	0,0830301

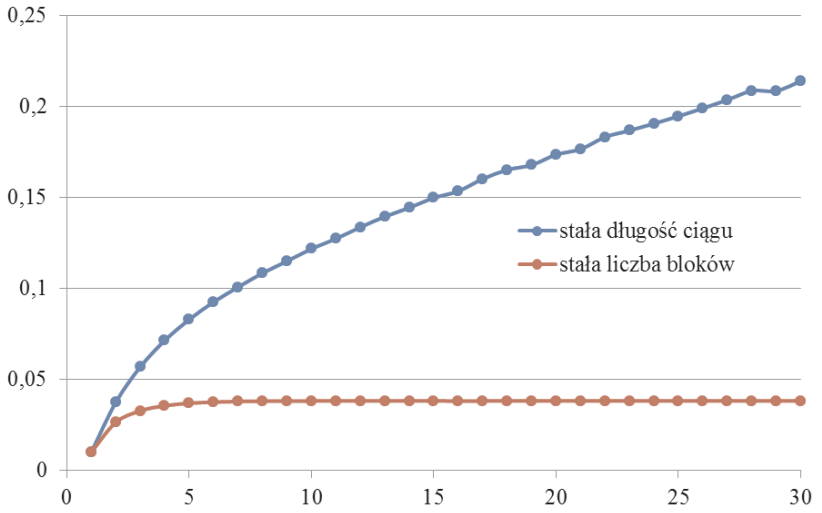
$d = 10$	
i	$n = 59$
0	0,0002394
1	0,0004787
2	0,0009575
3	0,0019149
4	0,0038299
5	0,0076596
6	0,0153181
7	0,0306278
8	0,0611880
9	0,1218381

$d = 20$	
i	$n = 29$
0	0,00000033
1	0,00000067
2	0,00000134
3	0,00000268
4	0,00000536
5	0,00001072
6	0,00002144
7	0,00004288
8	0,00008576
9	0,00017151
10	0,00034303
11	0,00068606
12	0,00137212
13	0,00274423
14	0,00548841
15	0,01097643
16	0,02194976
17	0,04387475
18	0,08755192
19	0,17354410

$d = 30$	
i	$n = 19$
0	0,0000000004
1	0,0000000008
2	0,0000000016
3	0.0000000032
4	0,0000000065
5	0,0000000130
6	0,0000000260
7	0,0000000519
8	0,0000001039
9	0,0000002078
10	0,0000004155
11	0,0000008311
12	0,0000016621
13	0,0000033243
14	0,0000066486
15	0,0000132971
16	0,0000265942
17	0,0000531885
18	0,0001063770
19	0,0002127539
20	0,0004255078
21	0,0008510154
22	0,0017020290
23	0,0034040470
24	0,0068080010
25	0,0136152700
26	0,0272246400
27	0,0544021700
28	0,1084294000
29	0,2139194000

Niebieski kolor odpowiada sytuacji przedstawionej w tabeli 6. Warto zauważyć, że począwszy od $d = 8$, obserwowany przyrost, można traktować jako liniowy – aproksymacja liniowa ma dla tego fragmentu współczynnik

dopasowania $R^2 = 0{,}9931$, a więc bardzo wysoki. Z kolei dla przypadku, gdy stała jest liczba bloków, począwszy od $d = 10$, wartość wpływu pozostaje praktycznie stała – łączny przyrost nie przekracza 0,2%. Wynika z tego, że dla większych długości bloku, przy stałej długości ciągu, za wzrost znaczenia najbardziej znaczącego bitu, odpowiada malejąca liczba bloków, jakie z takiego ciągu można wyodrębnić.



Rysunek 12. Zależności przeciętnego wpływu najbardziej znaczącego bitu bloku na wartość prawdopodobieństwa w funkcji długości bloku

Na podstawie obserwacji poczynionych w tym punkcie rekomendujemy, by zaproponowany test stosować dla podziału na bloki nie dłuższe niż 8 do 10 bitów. Jednocześnie przestrzegamy przed wykorzystaniem testu omówionego w punkcie 1.2. z użyciem tak krótkich bloków.

Dla obu tych testów, w przypadkach gdy konieczne byłoby przyjęcie dużych wartości d, obroną przed nieuwzględnianiem mniej znaczących bitów jest trik, znany z pakietu DIEHARD, polegający na serii podziałów przesuniętych względem siebie o pojedyncze pozycje bitowe. Trzeba jednak pamiętać, że prowadzi to do uzyskiwania serii skorelowanych statystyk, co może znacząco utrudnić interpretację otrzymywanych wyników.

4.2. Jest test dla ciągu bitów

Drugim spostrzeżeniem jest bliskość testu dla ciągu bitów i proponowanego testu przy jednobitowym bloku. Różnica między nimi sprowadza się

do zmiany operacji bitowej alternatywy wykluczającej (popularnie XOR) na iloczyn. Okazuje się ona jednak ważna i prowadzi do uzyskania dwóch istotnie różnych testów.

Przeprowadzone przez nas eksperymenty wskazują, że uzyskiwane dla obu testów, dla tego samego ciągu, prawdopodobieństwa wykazują dodatnią korelację na poziomie kilkunastu procent, a więc bardzo niskim. Co więcej punkty, których współrzędne odpowiadają obu testom, dość równomiernie pokrywają kwadrat jednostkowy.

Powyższe sprawia, że wyniki testu omówionego w punkcie 2.1. można traktować jako praktycznie niezależne od uzyskiwanych w teście proponowanym, w związku z czym wskazane i całkowicie poprawne jest wykorzystanie ich obok siebie.

Podsumowanie

Zaprezentowany w niniejszej pracy wariant testu autokorelacyjnego może, w badaniu ciągów binarnych, z powodzeniem zastąpić obydwa testy zaprojektowane dla ciągu liczbowego. Jednocześnie, dzięki stosowaniu krótkich bloków, możliwe jest usunięcie marginalizacji wpływu znaczącej części bitów, pojawiającej się podczas stosowania długich bloków.

Przeprowadzone eksperymenty pokazały, że dokładność zaproponowanego przybliżenia rozkładu statystyki testowej jest bardzo dobra już dla bloków 5 bitowych, przy wszystkich długościach badanego ciągu, zaś nawet dla podziału jednobitowego wymaga ciągów o niewygórowanej długości.

Na koniec chcemy dodać, że wszystkie te zalety osiągamy kosztem niewielkiej komplikacji procedury obliczającej dystrybuantę rozkładu statystyki testowej – wyznaczenie kilku współczynników zależnych od parametrów testu, wartości wielomianu 8 stopnia w punkcie oraz funkcji eksponent.

Literatura

[1] T.W. Anderson, D.A. Darling, *Asymptotic Theory of Certain Goodness of Fit Criteria Based on Stochastic Processes*, Annals of Mathematical Statistics, 23, 2, 1952

[2] P. LEcuyer, R. Simard, *TestU01: A Software Library in ANSI C for Empirical Testing of Random Number Generators*, Universite de Montreal, 2007

[3] B. Gniedenko B., A.N. Kołmogorow, *Rozkłady graniczne sum zmiennych losowych niezależnych*, PWN, Warszawa, 1957

[4] S.W. Golomb, *Shift Register Sequences*, San Francisco, Holden-Day, 1967
[5] D.E. Knuth, *Sztuka programowania*, t.2. Algorytmy seminumeryczne, WNT, 2002
[6] K. Mańk, *Dokładne dystrybuanty statystyk w testach momentów 1 i 2 rzędu*, Biuletyn WAT, KryptologiaIV, Warszawa, 2004
[7] G. Marsaglia, *DIEHARD Battery of Tests of Randomness*, 1995
[8] A. Rukhin i inni, *A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications*, NIST Special Publication 800–22, 2010
[9] R. Wieczorkowski, R. Zieliński, *Komputerowe generatory liczb losowych*, WNT, Warszawa, 1997

AUTOCORRELATION TEST FOR BINARY STREAM

Abstract. In paper we present a variant of autocorrelation test designed for binary streams divided into blocks. An approximation of tests statistic distribution as well as analysis of its quality is also given. Finally consequences of usage of long blocks are shown.

Keywords: statistical test, randomness test, autocorrelation test