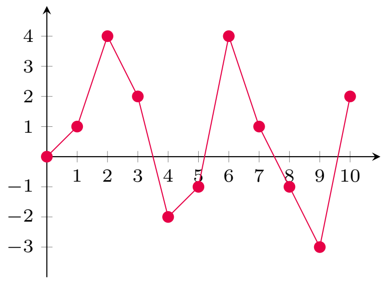
Twierdzenie Sparre–Andersena

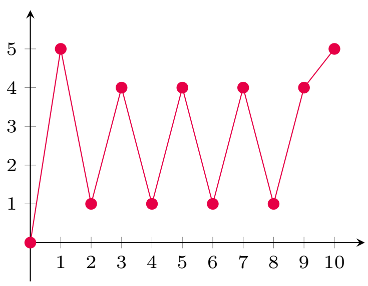
Łukasz RAJKOWSKI

W pewnej szkole w Symulandii każde z 1000 uczęszczających do niej dzieci dostało takie oto zadanie domowe: Trzeba rzucić 10 razy kostką sześcienną. W zależności od wyniku dziecko ma oddać lub pobrać od swoich rodziców kwotę określoną przez następującą tabelkę (ujemne kwoty oznaczają obowiązek oddania pieniędzy rodzicom):

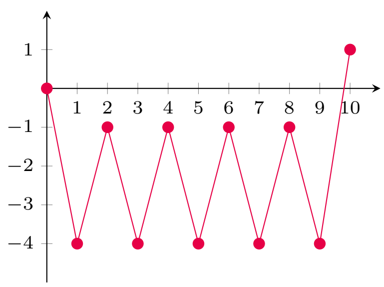
liczba oczek	**1**	**2**	**3**	**4**	**5**	**6**
wypłata	−4	−3	−2	1	3	5



Rys. 1. Przy takim przebiegu rozgrywki mamy $D = 6$ i $L = 2$



Rys. 2. Przy takim przebiegu rozgrywki mamy $D = 10$ i $L = 1$



Rys. 3. Przy takim przebiegu rozgrywki mamy $D = 1$ i $L = 10$

Następnie dzieci miały narysować wykres zależności stanu swojego konta od liczby rzutów (uwzględniając stan konta 0 w „zerowym rzucie") i na jego podstawie stwierdzić, ile rzutów kończyło się dodatnim stanem konta (co oznaczymy przez D) oraz po którym rzucie stan konta po raz pierwszy osiągnął największą wartość (co oznaczymy przez L). Rysunek 1 przedstawia przykładowy przebieg. Na pierwszy rzut oka wartości D i L nie są specjalnie związane – łatwo skonstruować przykłady, gdzie jedna z nich jest duża, a druga mała (rys. 2 i rys. 3). Gdy jednak dyrektor szkoły opublikował podsumowanie wyników pracy domowej, uwagę wszystkich zwróciła poniższa tabelka, przedstawiająca liczby uczniów, którzy uzyskali poszczególne wartości D (pierwszy wiersz) oraz poszczególne wartości L (drugi wiersz).

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
D	188	101	71	66	70	58	60	67	67	93	159
L	188	103	63	81	61	58	62	75	71	81	157

To, że tyle samo uczniów uzyskało wartości D i L równe 0, nie jest w ogóle zaskakujące – łatwo przekonać się, że byli to po prostu ci sami uczniowie. Dla pozostałych wartości D i L liczby uczniów, którzy je uzyskali, wydają się zaskakująco bliskie, choć przedstawione wcześniej przykłady pokazywały, że jeden uczeń mógł uzyskać istotnie różne wartości D i L. Dyrektor szkoły doniósł o swoim odkryciu premierowi Symulandii, który – poruszony tym nieoczekiwanym związkiem – zarządził ogólnokrajowy eksperyment. Każdy z 10^6 obywateli miał przeprowadzić to samo doświadczenie i przesłać uzyskane wyniki pocztą do Ministerstwa Liczenia. Pracujący tam urzędnicy skrupulatnie policzyli wszystkie wyniki i wyprodukowali tabelkę analogiczną do poprzedniej, z tą różnicą, że przedstawiała ona procentowy udział poszczególnych wyników w całej populacji, zaokrąglony do dziesiątych części procenta.

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
D	19,5%	10,3%	7,8%	6,6%	6,2%	6,0%	6,1%	6,4%	6,9%	8,3%	15,9%
L	19,5%	10,3%	7,8%	6,6%	6,2%	6,0%	6,1%	6,4%	6,9%	8,3%	15,9%

Kto jak kto, ale mieszkańcy Symulandii wiedzieli, że w tej sytuacji nie może być mowy o żadnym przypadku. Najwyraźniej prawdopodobieństwa uzyskania poszczególnych wartości przez zmienne D i L są równe! Zależność ta utrzymywała się w kolejnych eksperymentach, z różnymi wypłatami oraz kośćmi, które nie były dobrze wyważone. Symulandcy uczeni prędko znaleźli uzasadnienie tej własności, które oparte jest na pewnym kombinatorycznym twierdzeniu. Zanim je sformułujemy, wprowadzimy kilka oznaczeń.

Niech $\boldsymbol{x} = (x_1, \ldots, x_n)$ będzie dowolnym ciągiem liczb rzeczywistych. *Ustawieniem* $\boldsymbol{x}$ nazwiemy dowolny ciąg $(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$, gdzie σ jest pewną permutacją zbioru $\{1, 2 \ldots, n\}$. Subtelna różnica między ustawieniami a permutacjami polega na tym, że ponieważ wyrazy ciągu $\boldsymbol{x}$ mogą się powtarzać, ustawień tego ciągu może być mniej niż $n!$. W szczególności ciąg stały ma tylko jedno ustawienie. Niech $\boldsymbol{s} = (s_0, \ldots, s_n)$ będzie ciągiem sum częściowych $\boldsymbol{x}$, tzn. $s_0 = 0$ i $s_k = \sum_{i=1}^{k} x_i$ dla $1 \leqslant k \leqslant n$. Niech $D_{\boldsymbol{x}}$ oznacza liczbę dodatnich wyrazów w ciągu $\boldsymbol{s}$ oraz niech $L_{\boldsymbol{x}}$ będzie najmniejszym indeksem w ciągu $\boldsymbol{s}$, dla którego przyjmuje on największą wartość. Okazuje się, że dla każdego ciągu $\boldsymbol{x}$ możemy znaleźć jego ustawienie $\boldsymbol{x}'$ tak, aby $L_{\boldsymbol{x}} = D_{\boldsymbol{x}'}$.

Dla ciągu $\boldsymbol{x} = (1, 3, -2, -4, 1, 5, -3, -2, -2, 5)$ (którego sumy częściowe zostały zilustrowane na rysunku 1) kolejne kroki algorytmu przedstawia poniższa tabelka. W lewej kolumnie wypisana jest postać ciągu bazowego przed wykonaniem danego kroku, w środku suma tego ciągu bazowego, a w prawej kolumnie wartość dodawana na początku ciągu tworzonego po wykonaniu tego kroku.

ciąg bazowy	Σ	$+$
$1, 3, -2, -4, 1, 5, -3, -2, -2, 5$	2	1
$3, -2, -4, 1, 5, -3, -2, -2, 5$	1	3
$-2, -4, 1, 5, -3, -2, -2, 5$	-2	5
$-2, -4, 1, 5, -3, -2, -2$	-7	-2
$-2, -4, 1, 5, -3, -2$	-5	-2
$-2, -4, 1, 5, -3$	-3	-3
$-2, -4, 1, 5$	0	5
$-2, -4, 1$	-5	1
$-2, -4$	-6	-4
-2	-2	-2

W tej sytuacji

$\boldsymbol{x}' = (-2, -4, 1, 5, -3, -2, -2, 5, 3, 1)$.

Mamy wówczas

$\boldsymbol{s}' = (0, -2, -6, -5, 0, -3, -5, -7, -2, 1, 2)$,

zatem $D_{\boldsymbol{x}'} = 2 = L_{\boldsymbol{x}}$.

Twierdzenie. *Istnieje taka bijekcja $\Psi\colon \mathbb{R}^n \to \mathbb{R}^n$, że dla dowolnego $\boldsymbol{x} \in \mathbb{R}^n$ ciąg $\boldsymbol{x}' = \Psi(\boldsymbol{x})$ jest ustawieniem $\boldsymbol{x}$ oraz $L_{\boldsymbol{x}} = D_{\boldsymbol{x}'}$.*

Dowód. Wybierzmy dowolnie ciąg $\boldsymbol{x} = (x_1, \ldots, x_n)$. Ciąg $\boldsymbol{x}'$ tworzymy zgodnie z następującą procedurą. Na początku naszym „ciągiem bazowym” jest $\boldsymbol{x}$, a „ciągiem tworzonym” jest ciąg pusty. W każdym kroku procedury patrzymy na sumę wszystkich wyrazów obecnego ciągu bazowego – jeśli jest ona dodatnia, zabieramy z niego pierwszy wyraz, w przeciwnym wypadku wyraz ostatni. Zabrany wyraz wstawiamy na początek ciągu tworzonego. Dla przykładu, w przypadku przedstawionym na marginesie, w szóstym kroku ciągiem bazowym jest $(-2, -4, 1, 5, -3)$. Suma jego wyrazów to -3, zatem zabieramy z niego ostatni wyraz i dokładamy do ciągu tworzonego, którym po wykonaniu tego kroku jest $(-3, -2, -2, 5, 3, 1)$.

Postulowany w twierdzeniu ciąg $\boldsymbol{x}'$ to ciąg utworzony po n krokach opisanej procedury. Oczywiście $\boldsymbol{x}'$ jest ustawieniem ciągu $\boldsymbol{x}$. Niech $\boldsymbol{s}$ i $\boldsymbol{s}'$ będą ciągami sum częściowych ciągów odpowiednio $\boldsymbol{x}$ i $\boldsymbol{x}'$. Za pomocą prostej indukcji możemy uzasadnić, że suma wyrazów k-tego ciągu bazowego to s'_{n-k+1} (pamiętajmy, że wyrazy zabierane z ciągów bazowych przechodzą na początek ciągu tworzonego). W tej sytuacji liczba dodatnich wyrazów ciągu $\boldsymbol{s}'$ (czyli $D_{\boldsymbol{x}'}$) to liczba etapów konstrukcji, w których suma ciągu bazowego była dodatnia. Pokażemy teraz, że ta liczba jest równa $L_{\boldsymbol{x}}$.

Załóżmy, że $L_{\boldsymbol{x}} = l$, tzn. maksimum ciągu sum częściowych $\boldsymbol{s}$ pojawia się po raz pierwszy na indeksie l. Zwróćmy uwagę, że zmiany w ciągu bazowym mają „gąsienicowy” charakter, tzn. zawsze zabieramy któryś z dwóch jego krańcowych wyrazów. Uzasadnimy, że wyrazy x_i dla $i \leqslant l$ są zabierane z ciągu bazowego tylko „z lewej strony”, a wyrazy x_i dla $i > l$ tylko „z prawej strony”. Wynika to z poniższych dwóch obserwacji:

- jeśli ciąg bazowy zaczyna się od x_{l+1}, to jego suma jest niedodatnia, gdyż $\sum_{i=l+1}^{l+m} x_i = s_{l+m} - s_l \leqslant 0$, więc w takim wypadku będziemy już czerpać tylko „z prawej strony” ciągu bazowego.
- jeśli ciąg bazowy kończy się na x_l, to jego suma jest dodatnia, gdyż $\sum_{i=l-m}^{l} x_i = s_l - s_{l-m} > 0$, więc wtedy będziemy już czerpać tylko „z lewej strony” ciągu bazowego.

W tej sytuacji dokładnie l razy czerpiemy „z lewej strony” ciągu bazowego, zatem tyle razy ma on dodatnią sumę. Zgodnie z wcześniejszymi obserwacjami oznacza to, że $D_{\boldsymbol{x}'} = l$. Pozostaje uzasadnić, że przekształcenie Ψ faktycznie jest bijekcją, co wynika z istnienia „procedury odwrotnej”, której nietrudną konstrukcję pozostawiamy Czytelnikowi. $\square$

Jaki jest związek przedstawionego twierdzenia z niesamowitym zjawiskiem zaobserwowanym w Symulandii? Możemy w tym wypadku utożsamiać zdarzenia elementarne z ciągami wyników kolejnych gier. Jeśli $\boldsymbol{x} = (x_1, \ldots, x_{10})$ jest pewnym takim ciągiem, to każde ustawienie tego ciągu ma to samo prawdopodobieństwo wystąpienia. W związku z tym, na mocy naszego twierdzenia, zdarzenia elementarne ze zbioru $\{D = d\} = \{\boldsymbol{x}\colon D_{\boldsymbol{x}} = d\}$ (dla pewnej ustalonej wartości d) są połączone we wzajemnie jednoznaczne pary ze zdarzeniami elementarnymi (o tym samym prawdopodobieństwie) ze zbioru $\{L = d\} = \{\boldsymbol{x}\colon L_{\boldsymbol{x}} = d\}$. Oznacza to, że prawdopodobieństwa uzyskania poszczególnych wartości przez zmienne D i L faktycznie są równe, tak jak sugerowały to badania przeprowadzone w Symulandii.

Czytelnicy trochę bardziej obyci z rachunkiem prawdopodobieństwa bez trudu zauważą, że przedstawione rozumowanie może być zastosowane dla dowolnych zmiennych losowych (nie tylko tych o rozkładzie dyskretnym), a także, że dowodzi ono tak naprawdę równości rozkładów wektorów losowych (D, S) i (L, S), gdzie S jest sumą wszystkich wypłat. Obserwacja ta nosi nazwę twierdzenia Sparre–Andersena, które w pełnym brzmieniu przytaczamy poniżej.

Twierdzenie (Sparre–Andersen). *Niech $X_1, \ldots, X_n$ będą niezależnymi zmiennymi losowymi o tym samym rozkładzie. Niech $S_0 = 0$, $S_k = \sum_{i=1}^{k} X_i$ dla $1 \leqslant k \leqslant n$ oraz*

$$D = \sum_{i=1}^{n} \mathbf{1}_{\{S_i > 0\}}, \qquad L = \min\{0 \leqslant i \leqslant n\colon S_i = \max_{0 \leqslant j \leqslant n} S_j\}.$$

Wówczas wektory losowe (D, S) i (L, S) mają ten sam rozkład prawdopodobieństwa.