Liczenie ryb w jeziorze metodą statystyczną i śliczną, choć probabilistyczną

*Wydział Matematyki, Informatyki i Mechaniki, Uniwersytet Warszawski, Wydział Matematyki i Informatyki, Uniwersytet Mikołaja Kopernika w Toruniu

*Wojciech NIEMIRO**

W jeziorze pływa r ryb, ale liczby r nie znamy. Chcielibyśmy tę liczbę oszacować, nie uciekając się do osuszenia jeziora. Powiedzmy, że dysponujemy wędką, puszką farby i odrobiną wiedzy ze statystyki. Łowimy sobie jedną rybkę po drugiej i wrzucamy z powrotem do jeziora, krzywdy żadnej rybce nie czyniąc. Przed wrzuceniem do wody malujemy rybce kreseczkę na ogonku. Rybka złowiona powtórnie otrzymuje drugą kreseczkę. Jeśli zdarzy się złowić tę samą rybkę trzeci raz, domalowujemy trzecią kreseczkę i tak dalej. Wyniki naszych połowów zapisujemy w postaci ciągu $\mathbf{x} = (x_1, \ldots, x_n)$, gdzie x_i oznacza liczbę kresek na ogonku i-tej złowionej ryby przed wrzuceniem do jeziora. Jeśli, na przykład,

$$\mathbf{x} = (1,1,2,1,3,1,1,1,1,2,1,1,2,3,4,1,1,1,1,2,1,1,2,3,2),$$

to powtarzaliśmy połów 25 razy, złowiliśmy 15 różnych ryb, w tym jedną czterokrotnie, dwie trzykrotnie i trzy dwukrotnie. Jasne, że ciąg $\mathbf{x}$ zawiera pewną informację o nieznanej liczbie r. Duża liczba wyrazów ciągu różnych od jedynki (czyli ryb złowionych wielokrotnie) wskazuje na to, że r jest „prawdopodobnie małe”. Postaram się pokazać, jak to intuicyjne rozumowanie uściślić i sformułować wnioski w bardziej konkretnej, ilościowej postaci. Przy okazji zaprezentuję kilka ważnych idei, stojących u podstaw statystyki matematycznej.

Model probabilistyczny

Oczywiście, życie w jeziorze jest bardziej skomplikowane niż matematyka. Żeby coś obliczyć i przeprowadzić porządne rozumowanie, trzeba przyjąć szereg upraszczających założeń.

- Załóżmy, że liczba r jest niezmienna (ryby nie giną ani nie rozmnażają się).
- Pomiędzy kolejnymi połowami ryby całkowicie „mieszają się”. Mówiąc dokładniej, zakładamy, że w każdym kolejnym połowie prawdopodobieństwo wyciągnięcia każdej z ryb jest jednakowe, równe $1/r$.

Wyidealizowany model pozwala obliczyć prawdopodobieństwo otrzymania konkretnego wyniku połowu. Niech symbol $P_r(\mathbf{x})$ oznacza prawdopodobieństwo otrzymania wyniku $\mathbf{x}$ *przy założeniu*, że nieznana liczba ryb jest równa r. Dla przykładowych danych przytoczonych powyżej mamy

$$\begin{aligned} P_r(\mathbf{x}) &= P_r(1,1,2,1,3,1,1,1,1,2,1,1,2,3,4,1,1,1,1,2,1,1,2,3,2) \\ &= \frac{r}{r}\cdot\frac{r-1}{r}\cdot\frac{2}{r}\cdot\frac{r-2}{r}\cdot\frac{1}{r}\cdot\frac{r-3}{r}\cdot\frac{r-4}{r}\cdot\ldots\cdot\frac{11}{r}\cdot\frac{3}{r}\cdot\frac{10}{r} \\ &= \frac{r(r-1)\cdot\ldots\cdot(r-14)}{r^{25}}\cdot(2\cdot1\cdot6\cdot7\cdot2\cdot2\cdot10\cdot11\cdot3\cdot10) \\ &= \frac{(r)_{15}}{r^{25}}\cdot 1108800, \end{aligned}$$

gdzie użyliśmy oznaczenia $(r)_m = r(r-1)\cdot\ldots\cdot(r-m+1)$. Zauważmy, że 15 jest liczbą jedynek w ciągu $\mathbf{x}$ (liczbą różnych złowionych ryb). Łatwo wyjaśnić wyżej napisany wzór, przyglądając się kolejnym ułamkom w drugiej linii:

1. Pierwszy wyraz ciągu, x_1, zawsze *musi być* równy 1: na początku w jeziorze pływa r ryb i wszystkie są nieoznakowane. Pierwszy czynnik jest równy $r/r = 1$.
2. Po pierwszym połowie w jeziorze pływa $r-1$ ryb nieoznakowanych i jedna ryba oznaczona jedną kreską. Stąd prawdopodobieństwo otrzymania $x_2 = 1$ (wyłowienia nowej rybki) wynosi $(r-1)/r$.
3. Jeśli $x_1 = 1$ i $x_2 = 1$, to po drugim połowie w jeziorze pływa $r-2$ ryb nieoznakowanych i dwie ryby oznaczone jedną kreską. Prawdopodobieństwo otrzymania $x_3 = 2$ (wyłowienia oznakowanej rybki) wynosi więc $2/r$.

Rozwiązanie zadania M 1534. Niech $A_1A_2A_3A_4A_5A_6$ będzie sześciokątem foremnym o boku 1, a O będzie środkiem okręgu opisanego na tym sześciokącie. Wówczas

$$\mathcal{S} = \{O, A_1, A_2, A_3, A_4, A_5, A_6\}$$

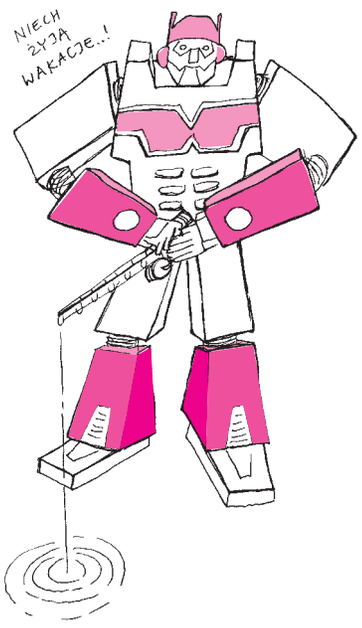
jest siedmioelementowym zbiorem spełniającym warunki zadania, gdyż punkt A_i jest środkiem okręgu opisanego na trójkącie $OA_{i-1}A_{i+1}$ dla $i = 1,2,\ldots,6$ (przyjmujemy $A_6 = A_0$ i $A_7 = A_1$).

Uwaga. Powyższy przykład można zastosować do konstrukcji zbioru n-elementowego spełniającego zadane warunki dla dowolnego $n \geqslant 7$. Czytelnika zachęcamy do rozstrzygnięcia, czy istnieje odpowiedni zbiór mocy 6 lub mniejszej.

Oznaczmy przez B_i środek okręgu opisanego na trójkącie OA_iA_{i+1} dla $i = 1,2,\ldots,6$, a przez τ – dowolną translację o wektor długości większej od 2. Wówczas, jeżeli $n = 7k + r$ dla $k \geqslant 1$ oraz $r \in \{0,1,\ldots,6\}$, to zbiór

$$\bigcup_{i=1}^{r}\{B_i\} \cup \bigcup_{\ell=0}^{k-1}\tau^\ell(\mathcal{S})$$

ma n elementów i spełnia warunki zadania (τ^ℓ oznacza ℓ-krotne złożenie τ).



4. Jeśli $x_1 = 1$, $x_2 = 1$ i $x_3 = 2$, to po trzecim połowie w jeziorze pływa $r - 2$ ryb nieoznakowanych. Prawdopodobieństwo otrzymania $x_4 = 1$ (wyłowienia jednej z tych nieoznakowanych) wynosi $(r - 2)/r$.
5. Jeśli $x_1 = 1$, $x_2 = 1$, $x_3 = 2$ i $x_4 = 1$, to po czwartym połowie w jeziorze pływa jedna ryba oznaczona dwiema kreskami. Prawdopodobieństwo otrzymania $x_5 = 3$ (wyłowienia właśnie tej rybki) wynosi $1/r$.

I tak dalej. Proponuję, żeby Czytelnik samodzielnie prześledził pochodzenie dalszych ułamków w naszym wzorze.

Wiarygodność

Wielkość $P_r(\mathbf{x})$ jest funkcją dwóch argumentów: $\mathbf{x}$ jest wynikiem doświadczenia losowego, a r jest nazywane parametrem. Możliwe są dwa punkty widzenia, charakteryzujące dwie różne dziedziny matematyki.

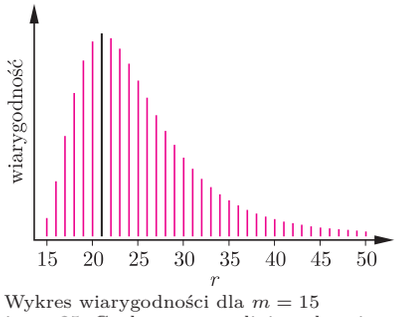
- Jeśli r jest ustalone (w domyśle – znane), to $P_r(\mathbf{x})$ rozważane *jako funkcja argumentu* $\mathbf{x}$ nazywa się *prawdopodobieństwem* (dokładniej – rozkładem prawdopodobieństwa). To jest punkt widzenia probabilistów.
- Jeśli $\mathbf{x}$ jest ustalone (w domyśle – znane), to $P_r(\mathbf{x})$ rozważane *jako funkcja argumentu* r nazywa się *wiarygodnością*. To jest punkt widzenia statystyków matematycznych.

W języku potocznym prawdopodobieństwo i wiarygodność są niemal synonimami, ale w naszych rozważaniach różnica między tymi pojęciami jest istotna. Zadanie, które postawiliśmy na początku tego artykułu: oszacowanie nieznanej liczby r na podstawie obserwacji $\mathbf{x}$ – należy do domeny statystyki.

Nasuwa się pomysł, że rozsądnym oszacowaniem parametru r jest taka wartość $\hat{r}$, która maksymalizuje wiarygodność

$$P_{\hat{r}}(\mathbf{x}) = \max_r P_r(\mathbf{x}).$$

Mówimy, że $\hat{r}$ jest *estymatorem największej wiarygodności* (ENW). Wróćmy do naszego przykładu. Dla ciągu $\mathbf{x}$, przytoczonego na początku artykułu, wiarygodność osiąga maksimum dla $r = 21$. Chciałoby się powiedzieć, że „21 jest najbardziej prawdopodobną liczbą ryb". Ale, ale! *Nie wolno* tak mówić! W naszym modelu r nie jest wynikiem jakiegoś doświadczenia losowego, a więc nie można mówić o „prawdopodobieństwie otrzymania r". Wobec tego statystycy mówią: „21 jest najbardziej *wiarygodną* liczbą ryb". Jest to wybieg językowy, który ukrywa dość zawiłą i niewygodną interpretację ENW. „Najbardziej prawdopodobny" po prostu znaczy „najczęściej pojawiający się w wielokrotnych powtórzeniach doświadczenia losowego". Ale co znaczy „najbardziej wiarygodny"?



Wykres wiarygodności dla $m = 15$ i $n = 25$. Grubsza czarna linia wskazuje ENW $\hat{r} = 21$.

- *ENW to jest taka wartość parametru, dla której, jeśliby wielokrotnie powtarzać doświadczenie losowe, to częściej otrzymywalibyśmy taki wynik, jaki w rzeczywistości otrzymaliśmy, w porównaniu z innymi możliwymi wartościami parametru.*

Dostateczność

Jak wynika z naszych dotychczasowych rozważań, wzór na wiarygodność w naszym rybackim zadaniu ma postać

$$P_r(\mathbf{x}) = P_r(x_1, \ldots, x_n) = \frac{(r)_m}{r^n} \cdot g(\mathbf{x}),$$

gdzie $m = m(\mathbf{x})$ jest *liczbą jedynek* w ciągu $\mathbf{x}$, zaś $g(\mathbf{x})$ jest funkcją $\mathbf{x}$, niezależącą od nieznanego r. Co prawda, ta funkcja jest raczej skomplikowana, ale nie będzie nam potrzebna! Zauważmy, że ENW możemy obliczyć, maksymalizując wyrażenie $P_r(\mathbf{x})$ z pominiętym czynnikiem $g(\mathbf{x})$. W rezultacie otrzymane oszacowanie $\hat{r} = \hat{r}(\mathbf{x})$ zależy tylko od $m = m(\mathbf{x})$. Okazuje się, że tylko m, liczba jedynek, zawiera informację o nieznanej liczbie r, wszystkie inne wielkości związane z wektorem $\mathbf{x} = (x_1, \ldots, x_n)$ są nieistotne! Mówimy, że $m = m(\mathbf{x})$ jest *statystyką dostateczną*. Następujące piękne rozumowanie przekona nas, że tak jest naprawdę. Ponieważ mamy

$$P(\mathbf{x}|r) = L(r, m(\mathbf{x})) \cdot g(\mathbf{x})$$

dla pewnej funkcji $L(r,m)$, to zgodnie z definicją prawdopodobieństwa warunkowego

$$P_r(\mathbf{x}|m) = \frac{P_r(\mathbf{x})}{P_r(m)} = \frac{P_r(\mathbf{x})}{\sum_{\mathbf{x}':m(\mathbf{x}')=m} P_r(\mathbf{x}')} =$$
$$= \frac{L(r,m)\cdot g(\mathbf{x})}{\sum_{\mathbf{x}':m(\mathbf{x}')=m} L(r,m)\cdot g(\mathbf{x}')} = \frac{g(\mathbf{x})}{\sum_{\mathbf{x}':m(\mathbf{x}')=m} g(\mathbf{x}')} = P(\mathbf{x}|m).$$

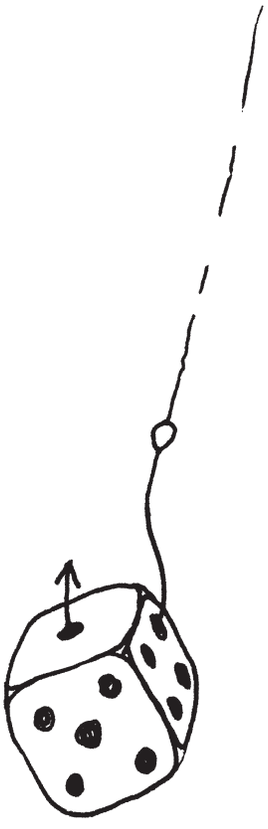
Prawdopodobieństwo warunkowe $P_r(\mathbf{x}|m)$ nie zależy od r, dlatego na końcu opuściliśmy indeks r. Przeprowadźmy następujące doświadczenie myślowe. Wyobraźmy sobie, że po dokonaniu połowu zapamiętaliśmy liczbę $m = m(\mathbf{x})$, a potem zgubiliśmy kartkę z zapisanym wektorem $\mathbf{x}$. Możemy odtworzyć zgubiony wynik doświadczenia, znając tylko m. Wylosujemy mianowicie fikcyjny wynik $\mathbf{x}'$ z prawdopodobieństwem $P(\mathbf{x}'|m)$, bo do tego nie jest potrzebna znajomość r. Chwila zastanowienia prowadzi do wniosku, że $\mathbf{x}'$ ma ten sam rozkład prawdopodobieństwa co $\mathbf{x}$. Skoro sposób naszego losowania nie zależał już od r, to uzyskany wynik nie mógł ze sobą nieść żadnej dodatkowej informacji o r. W tej sytuacji cała nasza wiedza o tym parametrze musi być „ukryta” w liczbie m!

Na zakończenie dodam, że tytuł tego artykułu zapożyczyłem z pięknego opowiadania Stanisława Lema *O królewiczu Ferrycym i królewnie Krystali* – opowieść z cyklu *Dzieła Cyfrotikon, czyli o dewijacyach, superfiksacyach, a waryacyach serdecznych.*

Jak zwalczać losowość w grach

*Bartłomiej ŻAK**

*Student, Wydział Matematyki, Informatyki i Mechaniki, Uniwersytet Warszawski



Losowość w grach karcianych, planszowych i komputerowych często budzi wiele kontrowersji. Sprawia ona, że gracz słabszy grający z lepszym ma szansę wygrać. Jest to pożądane w przypadku gier towarzyskich i frustrujące w przypadku gier profesjonalnych. W obu przypadkach nadmiar losowości jest zły, gdy za często zdarza się, że przewaga gracza pierwszego, wynikająca z jego inteligentnej gry, jest niwelowana przez szczęście drugiego. W moim artykule pokażę, jak z losowością można walczyć na przykładzie jednej z najbardziej losowych gier, czyli Chińczyka, którego, mam nadzieję, wszyscy znają.

Mój sposób mierzenia losowości jest zasadny dla gier z kategorii „wyścigów planszowych”. Takie gry w uproszczeniu polegają na tym, że kolejno losujemy liczby (na przykład rzucając kością) i chcemy, by suma naszych wylosowanych wartości jak najszybciej osiągnęła lub przekroczyła pewien pułap. Na przykład w Chińczyku, rzucając kośćmi ruszamy pionkiem o wylosowaną liczbę oczek, a żeby wygrać, musimy przesunąć pionki o 166 oczek (jeśli nie zbito naszego pionka). Te gry łączy jedno: nawet jeśli jesteś genialnym strategiem, jeśli będziesz miał pecha, to przegrasz.

W takich grach to, o ile zwiększy się nasza suma, warunkuje pewna zmienna losowa odpowiadająca jednemu losowaniu: będziemy o niej myśleć jak o obiekcie matematycznym, który przyjmuje różne wartości, każdą z pewnym prawdopodobieństwem, a prawdopodobieństwa te sumują się do jedynki. Dla przykładu, jeżeli zmienna losowa D reprezentuje wynik rzutu kością sześciościenną, to wartości tej zmiennej razem z prawdopodobieństwami wystąpienia wyglądają tak

x	1	2	3	4	5	6
$\mathbb{P}(D=x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Do opisu zmiennych losowych często używamy następujących pojęć: wartości oczekiwanej ($\mathbb{E}$) oraz wariancji (Var). Wartość oczekiwana to suma możliwych