O rybach i ufności

*Wojciech NIEMIRO**

*Zakład Statystyki Matematycznej, IMSM, WMIM, Uniwersytet Warszawski, Wydział Matematyki i Informatyki, Uniwersytet Mikołaja Kopernika w Toruniu

W poprzednim numerze *Delty* przedstawiliśmy zgrabną metodę szacowania liczby ryb pływających w stawie. Przypomnijmy doświadczenie, na którym ta metoda się opierała: najpierw łowimy rybkę, potem rysujemy jej kreskę na ogonku, następnie na kartce zapisujemy liczbę kresek, jakie widzimy na ogonku trzymanej w ręce rybki, po czym wrzucamy ją z powrotem do stawu i całą procedurę powtarzamy n razy.

Niech r będzie (nieznaną) liczbą ryb pływających w jeziorze. Poprzednio wykazaliśmy, że prawdopodobieństwo uzyskania na kartce konkretnego ciągu $\mathbf{x}$ wynosi $g(\mathbf{x})\frac{(r)_m}{r^n}$, gdzie m jest liczbą jedynek w tym ciągu (tzn. liczbą różnych, złowionych przez nas ryb), zaś $g(\mathbf{x})$ jest czynnikiem niezależnym od r. Wynika stąd, że m jest *statystyką dostateczną* i zawiera całą dostępną nam informację o r. Niech $P_r(m)$ oznacza prawdopodobieństwo wyłowienia dokładnie m różnych ryb. Nietrudno przekonać się, że $P_r(m) = \frac{(r)_m}{r^n}\left\{{n \atop m}\right\}$, gdzie $\left\{{n \atop m}\right\}$ jest liczbą podziałów zbioru n-elementowego na m rozłącznych podzbiorów (na tyle sposobów możemy złowić m różnych ryb przy n połowach).

Liczba podziałów zbioru n-elementowego na m rozłącznych podzbiorów nosi nazwę *liczby Stirlinga II rodzaju.* Można tę liczbę obliczyć przy użyciu wygodnej rekurencji. Prawdziwa jest zależność

$$\left\{{n \atop m}\right\} = \left\{{n-1 \atop m-1}\right\} + m\left\{{n-1 \atop m}\right\}.$$

Dlaczego?

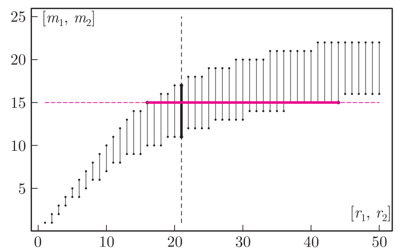
Wybierzmy teraz „małą” liczbę $\alpha > 0$ (na przykład $\alpha = 0{,}1$) i zdefiniujmy przedział $[m_1(r), m_2(r)]$ w następujący sposób:

$$m_1(r) = \text{ największa liczba } m_1, \text{ taka że } \sum_{m=1}^{m_1-1} P_r(m) \leqslant \alpha/2,$$

$$m_2(r) = \text{ najmniejsza liczba } m_2, \text{ taka że } \sum_{m=m_2+1}^{r} P_r(m) \leqslant \alpha/2.$$

Wynika stąd, że

$$(1)\qquad P_r\left(m_1(r) \leqslant m \leqslant m_2(r)\right) = \sum_{m=m_1(r)}^{m_2(r)} P_r(m) \geqslant 1-\alpha.$$



Konstrukcja przedziału ufności dla $m = 15$ i $n = 25$, na poziomie 90%. Pionowe linie są przedziałami o prawdopodobieństwie (co najmniej) 90%. Przedział dla $r = 21$ został wyróżniony tylko dla ułatwienia objaśnień. Poziomy odcinek jest przedziałem ufności.

Nierówność (1) mówi o tym, że z „dużym prawdopodobieństwem” $1-\alpha$ losowa wielkość m należy do przedziału $[m_1(r), m_2(r)]$, który zależy od nieznanego r. Na rysunku pionowe odcinki przedstawiają przedziały obliczone dla $\alpha = 0{,}1$ i różnych wartości r (od 1 do 50). Przykładowo, dla $r = 21$ mamy $m_1(r) = 11$, $m_2(r) = 17$ i $P_r(11 \leqslant m \leqslant 17) = 0{,}9600163$.

Przedstawione zależności wynikają z patrzenia na nasz rysunek *pionowo*, czyli dla różnych, ale ustalonych wartości r. To jest punkt widzenia probabilisty. Punkt widzenia statystyka jest *poziomy.* Rozpatrujemy ustaloną (bo zaobserwowaną) wartość m. Zdefiniujmy dwie zależne od m liczby „na osi poziomej”:

$$r_1(m) = \text{ najmniejsza liczba } r_1, \text{ taka że } m_2(r_1) \geqslant m,$$
$$r_2(m) = \text{ największa liczba } r_2, \text{ taka że } m_1(r_2) \leqslant m.$$

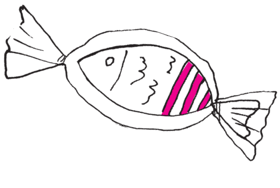
Na przykład, dla $m = 15$ mamy $r_1(m) = 16$ i $r_2(m) = 44$. Przedział $[16, 44]$ na „wysokości” $m = 15$ jest na rysunku 2 wyróżniony.

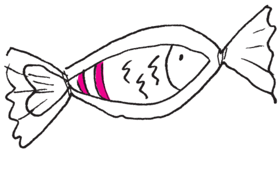
Doszliśmy teraz do najważniejszego miejsca naszych rozważań. Chwila zastanowienia prowadzi do wniosku, że następujące dwa warunki są równoważne:

$$r_1(m) \leqslant r \leqslant r_2(m) \qquad \text{oraz} \qquad m_1(r) \leqslant m \leqslant m_2(r).$$

W istocie, wynika to z definicji $r_i(m)$ i z faktu, że obie funkcje $m_i(r)$ są niemalejące, co nietrudno sprawdzić. Wynika stąd zatem, że dla każdego r

$$(2)\qquad P_r\left(r_1(m) \leqslant r \leqslant r_2(m)\right) \geqslant 1-\alpha.$$





Nierówność (2) mówi o tym, że dla dowolnego r, przedział $[r_1(m), r_2(m)]$ zawiera nieznaną liczbę r z dużym prawdopodobieństwem. Ten przedział możemy obliczyć, bo znamy m. Wspaniale! Wróćmy do naszych przykładowych danych, które pojawiły się na początku artykułu. Dla $m = 15$ (i ustalonego $n = 25$), przypomnijmy, $[r_1(m), r_2(m)] = [16, 44]$. A więc wydaje się, że następujące stwierdzenie jest zgodne z tym, co było powiedziane.

): *Przedział* $[16, 44]$ *zawiera nieznaną liczbę* r *z prawdopodobieństwem co najmniej* 0,90.

Równie bezsensowne jest stwierdzenie „przedział
[3,14159265358979323846l,
3,14159265358979323846 2]
zawiera liczbę π z prawdopodobieństwem co najmniej 0,90". Albo zawiera, albo nie. Chwilowo mogę nie wiedzieć, która z alternatywnych możliwości zachodzi, ale o żadnym prawdopodobieństwie nie można mówić! Jak się zajrzy do Wikipedii, to się wyjaśni.

Ale, ale, chyba się zagalopowaliśmy. Jeśli liczba r nie jest zmienną losową, to powyższe zdanie jest *bezsensowne*. Przedział $[16, 44]$ albo zawiera r, albo nie. Jak się jezioro osuszy, to się wyjaśni. Bez osuszania jeziora musimy nasz wniosek sformułować inaczej.

(: *Przedział* $[16, 44]$ *jest przedziałem ufności dla nieznanej liczby* r *na poziomie ufności* 0,90.

Jeśli o prawdopodobieństwie nie możemy mówić, to zastępujemy termin „prawdopodobieństwo" terminem „ufność". Matematyczną definicją przedziału ufności jest nierówność (2). Kłopot w tym, że prawdopodobieństwo we wzorze (2) opisuje niepewność wyniku doświadczenia, w tym przypadku wyłowienia m różnych ryb, przed wykonaniem doświadczenia (przed połowem). Jak więc interpretować przedział $[16, 44]$ obliczony *po* wyłowieniu $m = 15$ ryb?

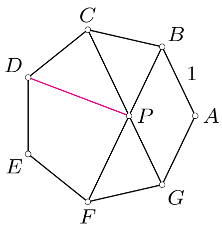
• *Przedział ufności na poziomie* $1 - \alpha$ *jest to przedział obliczony na podstawie wyniku doświadczenia losowego w taki sposób, że jeśliby powtarzać doświadczenie wielokrotnie, to dla przynajmniej* $(1 - \alpha) \cdot 100\%$ *doświadczeń, przedział obliczony tą samą metodą zawierałby nieznany parametr.*

W języku potocznym – „gdybanie".

Zwróćmy uwagę, jaką rolę w interpretacji przedziału ufności odgrywają zdania warunkowe i tryb przypuszczający. Jest to charakterystyczny dla Statystyka sposób myślenia – po wykonaniu doświadczenia losowego zastanawia się on: „z jakim prawdopodobieństwem to czy tamto by się mogło zdarzyć, gdyby nie to, że już się zdarzyło".

Zadania

Redaguje Łukasz BOŻYK



M 1537. Dany jest siedmiokąt foremny $ABCDEFG$ o boku długości 1. Przekątne BF i CG przecinają się w punkcie P. Znaleźć długość odcinka PD.
Rozwiązanie na str. 13

M 1538. Niech $n \geqslant 2$ będzie liczbą całkowitą. Znaleźć liczbę przedstawień liczby n w postaci sumy pewnej liczby dodatnich całkowitych składników, pośród których jest parzysta liczba liczb parzystych.
Rozwiązanie na str. 9

M 1539. Dana jest liczba $n \geqslant 1$ oraz pewien zbiór $A = \{a_1, a_2, \ldots, a_n\}$ dodatnich liczb całkowitych. Na okręgu wyróżniono 2^n punktów i każdemu z nich przyporządkowano jedną z liczb ze zbioru A. Udowodnić, że iloczyn liczb znajdujących się na pewnym łuku tego okręgu jest kwadratem liczby całkowitej.
Rozwiązanie na str. **??**

Przygotował Michał NAWROCKI

F 933. Pewien polaryzator przepuszcza $k_1 = 30\%$ padającej na niego wiązki niespolaryzowanego światła, a dwa takie polaryzatory, ustawione jeden za drugim, przepuszczają $k_2 = 13{,}5\%$ światła. Ile wynosi kąt α między płaszczyznami polaryzacji tych polaryzatorów?
Rozwiązanie na str. 1

F 934. Ile wynosi w przybliżeniu liczba cząsteczek powietrza zawartych w atmosferze ziemskiej? Przyjąć, że średnie ciśnienie atmosferyczne na powierzchni Ziemi wynosi 1013 hPa, średni promień Ziemi wynosi 6400 km, średnia masa cząsteczkowa powietrza (azot i tlen) wynosi $\mu = 29$ g/mol. Skorzystać z informacji, że satelita krążący wokół Ziemi na wysokości 100 km praktycznie nie napotyka oporu powietrza.
Rozwiązanie na str. 9