O twierdzeniach granicznych

Paweł STRZELECKI

Los każdego z nas zależy często od kolegialnych decyzji podejmowanych w głosowaniu przez najrozmaitsze ciała. Dlatego nie od rzeczy będzie rozważyć następujący

Problem. Parlament złożony z $n = 2m + 1$ osób podejmuje decyzję w drodze głosowania (zwykłą większością głosów). Załóżmy dla wygody, że istnieje obiektywnie określony *dobry wybór decyzji*. Przypuśćmy, że każdy z posłów w parlamencie rozpoznaje ów dobry wybór z prawdopodobieństwem $p = 0{,}7$, a przeciw dobremu wyborowi głosuje z prawdopodobieństwem $q = 1 - p = 0{,}3$. Przyjmijmy wreszcie, że każdy z posłów podejmuje swą decyzję w sposób niezależny. Pytanie brzmi: jaka jest najmniejsza liczba n posłów, dla której prawdopodobieństwo podjęcia przez ów parlament słusznej decyzji jest większe od 0,99?

Wielu licealistom problem wyda się może banalny. Mamy n prób Bernoulliego z prawdopodobieństwem sukcesu $p = 0{,}7$ i chodzi o to, by liczba sukcesów S_n była większa niż liczba porażek. Należy zatem sprawdzić, dla jakiego nieparzystego n zachodzi nierówność $P(S_n \geq m+1) > 0{,}99$ albo równoważnie $P(S_n \leq m) < 0{,}01$. Cóż prostszego, gdy jest jawny wzór,

Nierówność Czebyszewa stwierdza, że dla dowolnej zmiennej losowej X i dla $\varepsilon > 0$ mamy

$$P(|X - EX| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} D^2 X .$$

Gdy $X = S_n$ jest liczbą sukcesów w schemacie n prób Bernoulliego, to wartość oczekiwana i wariancja dane są dobrze znanymi wzorami

$$ES_n = np, \qquad D^2 S_n = np(1-p) .$$

W naszym przypadku, prawdopodobieństwo $P(S_n \leq m)$ nie przekracza lewej strony nierówności Czebyszewa, gdy położymy $\varepsilon = ES_n - m$. Prosty rachunek pokazuje, że dla takiej wartości ε wyrażenie $D^2 S_n/\varepsilon^2$ jest mniejsze od 0,01 dla $m \geq 260$, albo równoważnie dla $n \geq 521$.

$$(1) \qquad P(S_n \leq m) = \sum_{k=0}^{m} P(S_n = k) = \sum_{k=0}^{m} \binom{n}{k} p^k (1-p)^{n-k} .$$

Kto chce, niech włączy komputer, a dostanie odpowiedź w kilka sekund – posłów w parlamencie powinno być przynajmniej 31 (29 to jeszcze za mało).

Wyobraźmy sobie jednak, że mamy do dyspozycji tylko kartkę i ołówek, no i może jeszcze dość prymitywny kalkulator (taki, co to pomoże wykonać cztery działania i jeszcze wyciągnąć pierwiastek kwadratowy). Obliczanie długiej sumy, a nawet któregokolwiek z jej składników „na piechotę" raczej odpada: to dobra metoda na to, by nabawić się odcisków od naciskania guzików kalkulatora. Narzędzia w rodzaju nierówności Czebyszewa też dają wynik bardzo niedoskonały. Spróbujmy więc możliwie prosto, ale i możliwie dokładnie, wyrazić sumę (1) w sposób przybliżony.

Wystartujemy z oczywistej równości

$$P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k} = P(S_n = k+1) \cdot \frac{1-p}{p} \cdot \frac{k+1}{n-k} .$$

Korzystając z niej wielokrotnie, sprawdzamy, że

$$\begin{aligned} P(S_n \leq m) &= P(S_n{=}m) + P(S_n{=}m{-}1) + P(S_n{=}m-2) + \cdots + P(S_n{=}0) = \\ &= P(S_n{=}m)\underbrace{\left(1 + \frac{(1-p)m}{p(n-m+1)} + \frac{(1-p)^2 m(m-1)}{p^2(n-m+1)(n-m+2)} + \cdots\right)}_{m+1 \text{ składników}} . \end{aligned}$$

Zastępując sumę w nawiasie sumą postępu geometrycznego zostawiamy bez zmian jej dwa największe składniki. W trzecim składniku w miejsce ułamka $\frac{m-1}{n-m+2}$ wpisujemy $\frac{m}{n-m+1}$. Można sprawdzić, że w ten sposób zwiększamy wartość całej sumy o mniej, niż $\frac{2}{5m}$. Dociekliwy Czytelnik zechce samodzielnie ocenić, jaki błąd popełniamy w przypadku kolejnych składników.

Dokonajmy teraz podwójnego drobnego oszustwa. Po pierwsze, nie popełniając wielkiego błedu, można przyjąć, że składniki sumy w nawiasie to kolejne potęgi tej samej liczby $x = \frac{1-p}{p} \cdot \frac{m}{n-m+1}$. Stąd

$$P(S_n \leq m) \approx P(S_n = m)(1 + x + x^2 + \cdots + x^m) .$$

Drugie uproszczenie polega na tym, by sumę skończoną zastąpić sumą nieskończoną. Zysk jest spory, bo sumę nieskończonego ciągu geometrycznego oblicza się łatwiej niż skończonego, a błąd niewielki: nietrudno sprawdzić, że w naszym przypadku $x \in (0, (3/7))$, więc już dla $m = 10$ suma $x^{m+1} + x^{m+2} + \cdots$ jest ponad 1000 razy mniejsza od pierwszego, równego jeden, składnika w nawiasie. W efekcie, po nietrudnym rachunku, dostaniemy

$$P(S_n \leq m) \approx \frac{P(S_n = m)}{1-x} = P(S_n = m)\frac{(n+1)p - mp}{(n+1)p - m} \approx \frac{p}{2p-1} P(S_n = m) .$$

Tym samym sprowadziliśmy problem obliczenia wartości długiej sumy do obliczenia największego składnika i pomnożenia go przez liczbę $p/(2p-1)$. W dalszych rachunkach pomoże nam

Lokalne twierdzenie graniczne. *Załóżmy, że* $0 < m < n$ *i oznaczmy przez* p^* *ułamek* $\frac{m}{n}$. *Jeśli* $m, n \to \infty$ *w ten sposób, że również* $n - m \to \infty$, *to zachodzi wtedy przybliżony wzór*

$$P(S_n = m) = \binom{n}{m} p^m (1-p)^{n-m} \simeq \frac{1}{\sqrt{2\pi n p^*(1-p^*)}} \cdot \exp(-nH(p^*)), \tag{2}$$

gdzie $H(x) = x \ln \frac{x}{p} + (1-x) \ln \frac{1-x}{1-p}$.

Znak $\simeq$ w równości (2) należy rozumieć w ten sposób, że stosunek lewej strony do prawej dąży do jedności dla $m, n \to \infty$.

Szkic dowodu. Lokalne twierdzenie graniczne jest prostą konsekwencją wzoru Stirlinga, pozwalającego na przybliżone obliczanie wartości silni, $n! \simeq n^n e^{-n}\sqrt{2\pi n}$. Skorzystajmy z tego wzoru, by wyrazić wszystkie silnie występujące w definicji prawdopodobieństwa $P(S_n = m)$. Dostaniemy w efekcie równość przybliżoną

Dokładniej, wzór Stirlinga (pochodzący z lat trzydziestych XVIII wieku) mówi, że

$$n! = n^n e^{-n+\rho(n)}\sqrt{2\pi n},$$

gdzie $\rho(n)$ jest liczbą z przedziału $(\frac{1}{12n+1}, \frac{1}{12n})$.

$$P(S_n = m) \simeq \sqrt{2\pi m(n-m)} \cdot \frac{n^n}{m^m (n-m)^{n-m}} \cdot p^m (1-p)^{n-m} =$$
$$= \frac{1}{\sqrt{2\pi n p^*(1-p^*)}} \cdot \left(\frac{p}{p^*}\right)^m \cdot \left(\frac{1-p}{1-p^*}\right)^{n-m}.$$

Stąd już każdy bez kłopotu dowędruje do tezy twierdzenia: wystarczy posłużyć się zależnością $a = e^{\ln a}$ i wiedzieć, że logarytm iloczynu dwóch liczb to suma ich logarytmów.

Może się wydawać, że lokalne twierdzenie graniczne oferuje wzór bardzo zawiły zamiast względnie prostego i elementarnego. To jednak tylko pozór, bowiem do obliczeń dla dużych wartości n znacznie wygodniej jest użyć formuły (2) niż definicji rozkładu Bernoulliego.

Jeśli funkcja f ma ciągłe pochodne do rzędu k włącznie, to jej wartości w pobliżu ustalonego punktu x_0 można przybliżać korzystając ze wzoru Taylora,

$$f(x_0+h) = f(x_0) + f'(x_0)h + \frac{f''(x_0)}{2!}h^2 + \dots + \frac{f^{(k)}(x_0)}{k!}h^k + R(x_0, h).$$

Reszta $R(x_0, h)$ spełnia warunek

$$\lim_{h\to 0} \frac{R(x_0,h)}{h^k} = 0.$$

O ile f ma ciągłą pochodną rzędu $(k+1)$, to istnieje taka stała C, że dla małych h mamy $|R(x_0,h)| < C|h|^{k+1}$. Do przybliżania wartości funkcji H stosujemy obok wzór Taylora dla $k = 1$, a nieco dalej – dla $k = 2$.

W naszym przypadku liczba $p^* = \frac{m}{n} = \frac{1}{2} - \frac{1}{2n}$ niewiele się różni od $\frac{1}{2}$. Zamiast więc obliczać $H(p^*)$ bezpośrednio, wypiszemy równanie stycznej do wykresu funkcji H w punkcie $\left(\frac{1}{2}, H\left(\frac{1}{2}\right)\right)$ i odczytamy z owego równania przybliżoną wartość $H(p^*)$. Ponieważ współczynnik kierunkowy stycznej to pochodna $H'\left(\frac{1}{2}\right)$, mamy

$$H(p^*) \approx H\left(\frac{1}{2}\right) + H'\left(\frac{1}{2}\right)\left(p^* - \frac{1}{2}\right) = -\ln\sqrt{4p(1-p)} - \frac{1}{n}\ln\sqrt{\frac{1-p}{p}}.$$

Podstawiając tę wartość do wzoru (2), i jeszcze raz wykorzystując fakt, że $p^* \approx \frac{1}{2}$, dostajemy

$$P(S_n \le m) \approx \frac{p}{2p-1} P(S_n = m) \approx \frac{p}{2p-1}\sqrt{\frac{2}{\pi n}} \cdot \exp\left(n \ln\sqrt{4p(1-p)} + \ln\sqrt{\frac{1-p}{p}}\right) =$$
$$= \sqrt{\frac{2p(1-p)}{\pi n}} \cdot \frac{(4p(1-p))^{n/2}}{2p-1}.$$

Posługując się kalkulatorem otrzymamy

$$P(S_n \le m) \approx 0{,}914\ldots \cdot \frac{1}{\sqrt{n}} \cdot (0{,}916\ldots)^n \underset{\text{ozn.}}{=} a(n).$$

Wrogowie stosowania kalkulatorów i miłośnicy spokojnych rachunków, którzy nie boją się wypisać wzoru Taylora dla funkcji $\ln(1+x)$ i wiedzą, że $e = 2{,}7182\ldots$, mogą poszukać rozwiązań nierówności $a(n) < 0{,}01$ korzystając jedynie z ołówka i kartki papieru (to wcale nie takie straszne, jak by się mogło wydawać).

Jest to już przybliżenie nadające się do rachunków na kalkulatorze. Spośród wszystkich liczb nieparzystych n najbliżej liczby rzeczywistej x, spełniającej zależność $a(x) = 0{,}01$, leży właśnie $n = 31$. Lawina drobnych „oszustw" nie była więc wcale niebezpieczna.

Z naszych rachunków można wywnioskować więcej. Mianowicie, jeśli liczba prób n jest duża, to powinniśmy *oczekiwać* uzyskania około $E(S_n) = np$ sukcesów. W takim przypadku $p^* = m/n$ niewiele się różni od p. Obliczmy więc, jaka jest przybliżona wartość funkcji $H(p^*)$ dla $p^* \approx p$. Jak poprzednio, użyjemy wzoru Taylora – tym razem biorąc o jeden wyraz więcej, bo tyle trzeba wziąć, żeby uzyskać nietrywialny wynik. Mamy $H(p) = H'(p) = 0$, zaś

$H''(p) = \frac{1}{p(1-p)}$. Zatem

$$H(p^*) = \frac{1}{2p(1-p)}(p^* - p)^2 + R(p^*)\,,$$

przy czym reszta R spełnia warunek $|R(p^*)| \le C(p^* - p)^3$ dla p^* bliskich p. Podstawiając tę wartość do wzoru (2), dostaniemy

$$P(S_n = m) \approx \frac{1}{\sqrt{2\pi np(1-p)}} \cdot \exp\left(-\frac{n(p^* - p)^2}{2p(1-p)}\right)$$

albo, oznaczając $z = n(p^* - p)$, $\delta = 1/\sqrt{np(1-p)}$,

$$P(S_n = m) \approx \delta\phi(\delta z), \qquad \text{gdzie} \quad \phi(t) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2}.$$

Jeśli funkcja f jest (np.) ciągła na przedziale $[a, b]$, to tzw. *sumy całkowe Riemanna*,

$$\sum_{k=1}^{n} \frac{b-a}{n} f\left(a + k\frac{b-a}{n}\right)$$

są zbieżne, dla $n \to \infty$, do *całki oznaczonej* $\int_a^b f(x)\,\mathrm{d}x$. Ten fakt ma interpretację geometryczną: pole pod wykresem (odpowiednio porządnej) funkcji można z dowolną dokładnością przybliżać sumą pól cieniutkich prostokątów.

Co to znaczy? Interpretacja nie jest trudna. Jeśli prób jest wiele, to interesują nas raczej zdarzenia $\{a \le S_n \le b\}$, a nie (znikomo prawdopodobne) zdarzenia $\{S_n = m\}$. Gdy skorzystamy z ostatniego wzoru, to dostaniemy równość przybliżoną

$$P(a \le S_n \le b) \approx \sum_{a-np \le z \le b-np} \delta\phi(\delta z)\,.$$

Suma po prawej stronie nie bez powodu przypomina sumę całkową Riemanna funkcji ϕ – Czytelnik sam zechce sprawdzić, po jakim przedziale odbywa się „całkowanie".

Dotarliśmy do końca – w tym momencie wypada sformułować

Twierdzenie graniczne de Moivre'a–Laplace'a. *Jeśli S_n oznacza liczbę sukcesów w n próbach Bernoulliego z prawdopodobieństwem sukcesu p, to*

$$\lim_{n\to\infty} P\left(\alpha \le \frac{S_n - np}{\sqrt{np(1-p)}} \le \beta\right) = \frac{1}{\sqrt{2\pi}} \int_\alpha^\beta \mathrm{e}^{-t^2/2}\,\mathrm{d}t$$

dla dowolnych $\alpha, \beta \in \mathbf{R}$, $\alpha < \beta$.

Twierdzenie de Moivre'a–Laplace'a ma dziś całą masę uogólnień o dużym znaczeniu tak teoretycznym, jak i praktycznym. Rachunkowi prawdopodobieństwa trudno byłoby się obyć bez tego rezultatu.

Ci Czytelnicy, którzy znają wzór Taylora oraz definicję całki Riemanna, nie powinni mieć kłopotu z zamienieniem treści niniejszego artykułu na ścisły i nie nazbyt długi dowód. Można będzie wtedy powiedzieć: „Rzeczywiście, nietrudne – nic dziwnego, że Laplace umiał to zrobić ponad 180 lat temu."

Rozwiązanie zadania M 760. Wykonajmy rzut stereograficzny naszej sfery z punktu P na płaszczyznę i oznaczmy obrazy punktów A, B, C, D, E i F odpowiednio przez A', B', C', D', E' i F'. Punkty A', D' i B' są współliniowe, podobnie – punkty B', E' i C' oraz punkty A', F' i C'. Obrazami okręgów opisanych w treści zadania są okręgi (jest to znana własność rzutu stereograficznego). Na mocy zadania **M 759** mają one punkt wspólny. Odwrotny rzut stereograficzny tego punktu (z płaszczyzny na sferę) jest szukanym punktem wspólnym okręgów leżących na sferze.