Prawo małych liczb

Rafał SZTENCEL *

Jeśli rzucamy n razy monetą, na której orzeł wypada z prawdopodobieństwem p, to szansa uzyskania dokładnie k orłów jest równa

$$p_k = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots n.$$

Innymi słowy, liczba orłów jest zmienną losową S_n, $P(S_n = k) = p_k$, $k = 0, 1, 2, \ldots, n$. Powiemy, że S_n ma rozkład Bernoulliego z parametrami n, p.

Jeśli n i k są duże, obliczenie p_k może być kłopotliwe. Okazuje się jednak, że gdy iloczyn np jest nieduży, istnieje proste przybliżenie. Niech $np = \lambda$. Wtedy

$$p_k \approx \pi_k = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots$$

Nietrudno stwierdzić, że

$$\sum_{k=0}^{\infty} \pi_k = 1.$$

Zatem (nieujemne) liczby π_k definiują pewien rozkład prawdopodobieństwa, zwany rozkładem Poissona z parametrem λ. Będziemy go oznaczać $Pois(\lambda)$.

Często można przeczytać, że przybliżenie $p_k \approx \pi_k$ opiera się na następującym twierdzeniu granicznym:

Twierdzenie Poissona. *Jeśli $n \to \infty$, $p_n \to 0$, $np_n \to \lambda$, to dla ustalonego $k \in \{0, 1, 2, \ldots\}$*

$$\binom{n}{k} p_n^k (1-p_n)^{n-k} \to \frac{\lambda^k}{k!} e^{-\lambda}.$$

Jeśli n jest duże, to mamy do czynienia z dalekim wyrazem ciągu, którego granicą jest π_k, więc pewnie uwierzymy, że wyraz ten jest bliski π_k. W takim razie dlaczego napisaliśmy wcześniej, że np ma być „nieduże"? Wszystko to wygląda raczej na metodologię empiryczną, tym bardziej że twierdzenie Poissona nie zawiera żadnych informacji o szybkości zbieżności. Może dałoby się uzyskać jakieś informacje, analizując jego dowód?

Zachęcamy Czytelnika do samodzielnego przeprowadzenia dowodu (nie jest taki trudny) i sprawdzenia, czy da się elegancko oszacować $|p_k - \pi_k|$.

Twierdzenie Poissona – w sformułowaniu innym niż przytoczone – pojawiło się w pracy [2] z roku 1837, poświęconej tzw. prawdopodobieństwu sądowemu. Natomiast w pracy Bortkiewicza [1] z roku 1898 pt. „Prawo małych liczb" można znaleźć cztery przykłady pokazujące, że rozkład Poissona pojawia się, gdy mamy do czynienia ze zdarzeniami rzadkimi. U Bortkiewicza są to samobójstwa wśród dzieci w Prusach (60 wypadków w latach 1869–1893), samobójstwa wśród kobiet w ośmiu krajach niemieckich (po kilkadziesiąt wypadków w latach 1881–1894), śmiertelne wypadki przy pracy wśród członków jedenastu związków zawodowych (61 wypadków w latach 1886–1894), i wreszcie zgony na skutek kopnięcia przez konia w czternastu korpusach armii pruskiej (196 wypadków w latach 1875–1894).

*Instytut Matematyki Uniwersytetu Warszawskiego

Poniżej zamieszczamy oryginalną tabelę z pracy Bortkiewicza. Dane te są dość często cytowane w okrojonej postaci, mianowicie z wyłączeniem korpusu gwardii (G) oraz korpusów I, VI i XI, miały one bowiem nietypowy skład.

24 Zweites Kapitel. § 12.

	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94
G	—	2	2	1	—	—	1	1	—	3	—	2	1	—	—	1	—	1	—	1
I	—	—	—	2	—	3	—	2	—	—	—	1	1	1	—	2	—	3	1	—
II	—	—	—	2	—	2	—	—	1	1	—	—	2	1	1	—	—	2	—	—
III	—	—	—	1	1	1	2	—	2	—	—	—	1	—	1	2	1	—	—	—
IV	—	1	—	1	1	1	1	—	—	—	—	1	—	—	—	—	1	1	—	—
V	—	—	—	—	2	1	—	—	1	—	—	1	—	1	1	1	1	1	1	—
VI	—	—	1	—	2	—	—	1	2	—	1	1	3	1	1	1	—	3	—	—
VII	1	—	1	—	—	—	1	—	1	1	—	—	2	—	—	2	1	—	2	—
VIII	1	—	—	—	1	—	—	1	—	—	—	—	1	—	—	—	1	1	—	1
IX	—	—	—	—	—	2	1	1	1	—	2	1	1	—	1	2	—	1	—	—
X	—	—	1	1	—	1	—	2	—	2	—	—	—	—	2	1	3	—	1	1
XI	—	—	—	—	2	4	—	1	3	—	1	1	1	1	2	1	3	1	3	1
XIV	1	1	2	1	1	3	—	4	—	1	—	3	2	1	—	2	1	1	—	—
XV	—	1	—	—	—	—	—	1	—	1	1	—	—	—	2	2	—	—	—	—

Liczba wypadków w danym korpusie w ciągu roku powinna mieć rozkład Poissona. Oto argument: szansa na wypadek w krótkim przedziale czasu (np. jednej godziny) jest niedużą liczbą p. Rok ma $n = 8760$ godzin, można uwierzyć, że np. wypadek o 6:30 rano 3 kwietnia nie wpłynie na szanse wypadku między 7:00 a 8:00 dnia 31 października. Niezależność prowadzi do schematu Bernoulliego, a rozkład Bernoulliego przybliżamy rozkładem Poissona.

Dane z tabeli pozwalają na obliczenie średniej liczby wypadków w korpusie w ciągu roku: jest to $\frac{196}{280} = 0{,}7$. Ale wartość średnia zmiennej losowej o rozkładzie $Pois(\lambda)$ jest równa λ, skąd naturalny pomysł, żeby zbadać zgodność danych z rozkładem $Pois(0{,}7)$.

Dokładniej, rozkład Poissona przewiduje, że prawdopodobieństwo roku bez wypadku jest równe

$$\pi_0 = e^{-\lambda} = 0{,}49658...,$$

zatem w $n = 280$ eksperymentach można się spodziewać średnio $n\pi_0 = 139{,}04$ takich przypadków, a otrzymano 144. W kolejnej tabeli porównujemy teoretyczną prognozę liczby korpusów, gdzie miało miejsce $0, 1, 2, \ldots$ wypadków ($n\pi_k$) z danymi empirycznymi (n_k).

k	π_k	n_k	$n \cdot \pi_k$
0	0,4966	144	139,04
1	0,3476	91	97,33
2	0,1217	32	34,07
3	0,0284	11	7,95
4	0,0050	2	1,39
$\geqslant 5$	0,0007	0	0,20

Zgodność jest na pierwszy rzut oka niezła, tym bardziej że dopasowanie rozkładu Poissona metodą średniej nie musi być przecież najlepsze. Wybór metody i ocena jakości dopasowania to już domena statystyki.

Pokażemy teraz, jak uzyskać oszacowanie błędu w przybliżeniu Poissona. Udowodnimy

Twierdzenie. *Niech X_1, X_2, …, X_n będą niezależnymi zmiennymi losowymi o tym samym rozkładzie:*

$$P(X_i = 1) = p = 1 - P(X_i = 0),$$

$i = 1, 2, \ldots, n$. Niech $\lambda = np$, $\pi_k = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}$ dla $k = 0, 1, 2, \ldots$, $S_n = X_1 + X_2 + \ldots + X_n$.

Wtedy dla każdego zbioru $B \subset \{0, 1, 2, \ldots\}$ mamy

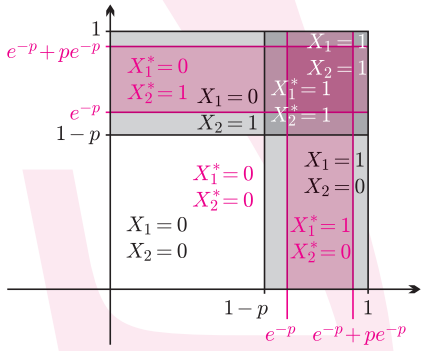
$$\left| P(S_n \in B) - \sum_{k \in B} \pi_k \right| \leqslant \frac{\lambda^2}{n}.$$

Mamy tu oszacowanie dotyczące nie tylko pojedynczych różnic $p_k - \pi_k$, ale prawdopodobieństw wszystkich możliwych zdarzeń!

Wynika stąd oczywiście twierdzenie Poissona (wystarczy wziąć $B = \{k\}$), a nawet więcej: zbieżność zachodzi, gdy $np_n \to \infty$, ale $np_n^2 \to 0$. Jak widać, uzasadniona jest zdroworozsądkowa reguła mówiąca, kiedy przybliżenie jest dobre: n ma być duże, p — małe, np — nieduże. Teraz jednak konkretne oszacowanie pozwala na podjęcie decyzji, czy przybliżenie uznajemy za wystarczająco dobre.

Warto zwrócić uwagę na dwa chwyty, których użyjemy w dowodzie. Po pierwsze, liczba sukcesów S_n w schemacie Bernoulliego jest sumą n niezależnych (i bardzo prostych) zmiennych losowych X_i, które liczą sukcesy w pojedynczych doświadczeniach. W dowodzie porównamy je ze zmiennymi losowymi X_i^*, które są niezależne i mają rozkład $Pois(p)$.

Po drugie, żeby zobaczyć, jak dalece X_i i X_i^* różnią się, wymodelujemy je na zbiorze zdarzeń elementarnych $\Omega = [0, 1]^n$, czyli kostce n-wymiarowej. Prawdopodobieństwo P jest n-wymiarowym odpowiednikiem objętości (szczerze mówiąc, miarą Lebesgue'a); przypadek $n = 2$ da się narysować.



Zmienne losowe X_1, X_2, X_1^*, X_2^* (Ω to $[0, 1] \times [0, 1]$).

To rzadki przykład, kiedy faktycznie zbiór Ω nie jest jedynie uciążliwym, choć niezbędnym, elementem teorii.

Być może Czytelnik domyśla się, że potrzebujemy jeszcze jednego faktu:

Lemat. *Suma niezależnych zmiennych losowych o rozkładzie $Pois(a)$ i $Pois(b)$ ma rozkład $Pois(a + b)$.*

Dowód lematu. Niech $k \in \{0, 1, 2, \ldots\}$ i niech X i Y będą zmiennymi losowymi, o których mowa w lemacie. Mamy

$$P(X + Y = k) = \sum_{j=0}^{k} P(X = j, Y = k - j) =$$

$$= \sum_{j=0}^{k} P(X = j)P(Y = k - j) = \sum_{j=0}^{k} \frac{a^j}{j!}\mathrm{e}^{-a} \cdot \frac{b^{k-j}}{(k-j)!}\mathrm{e}^{-b} =$$

$$= \sum_{j=0}^{k} \frac{a^j b^{k-j}}{j!(k-j)!}\mathrm{e}^{-(a+b)} = \frac{1}{k!}\mathrm{e}^{-(a+b)} \sum_{j=0}^{k} \binom{k}{j} a^j b^{k-j} =$$

$$= \frac{(a+b)^k}{k!}\mathrm{e}^{-(a+b)}.$$

Dowód twierdzenia. Określimy formalnie zmienne losowe X_k:

$$X_k(\omega) = X_k(\omega_1, \ldots, \omega_n) = \begin{cases} 0 & \text{dla } \omega_k < 1 - p, \\ 1 & \text{dla } \omega_k \geqslant 1 - p. \end{cases}$$

Są one niezależne. Dalej, niech

$$X_k^*(\omega) = \begin{cases} 0 & \text{dla } \omega_k < \mathrm{e}^{-p}, \\ k & \text{dla } \omega_k \in [a_{k-1}, a_k), \end{cases}$$

gdzie $a_k = \sum_{m=0}^{k} \pi_m$. Wtedy X_k^* są niezależne i na mocy lematu $S_n^* = X_1^* + \ldots + X_n^*$ ma rozkład Poissona z parametrem np.

W takim razie

$$\left| P(S_n \in B) - \sum_{k \in B} \pi_k \right| \leqslant$$

$$\leqslant |P(S_n \in B) - P(S_n^* \in B)| \leqslant P(S_n \neq S_n^*),$$

co wynika z elementarnej nierówności:

$$|P(A) - P(C)| \leqslant P(A \triangle C),$$

(gdzie $\triangle$ oznacza różnicę symetryczną zbiorów) i stąd, że zdarzenie $\{S_n \in B\} \triangle \{S_n^* \in B\}$, polegające na tym, że jeśli $S_n \in B$, to $S_n^* \notin B$ i odwrotnie, w oczywisty sposób pociąga za sobą, że $S_n \neq S_n^*$. Teraz

$$P(S_n \neq S_n^*) \leqslant P\left(\bigcup_{k=1}^{n} \{X_k \neq X_k^*\} \right) \leqslant \sum_{k=1}^{n} P(X_k \neq X_k^*).$$

Zmienne losowe X_k i X_k^* różnią się na dwóch zbiorach: $\{\omega\colon\ \omega_k \in [1 - p, \mathrm{e}^{-p})\}$ (tu $X_k = 0$, $X_k^* = 1$) oraz $\{\omega\colon\ \omega_k \in [\mathrm{e}^{-p} + p\mathrm{e}^{-p}, 1]\}$ (tu $X_k = 1$ i $X_k^* > 1$), których łączna miara nie przekracza p^2. Istotnie, ponieważ $\mathrm{e}^{-p} \geqslant 1 - p$, więc

$$\mathrm{e}^{-p} - (1 - p) + 1 - \mathrm{e}^{-p} - p\mathrm{e}^{-p} = p(1 - \mathrm{e}^{-p}) \leqslant p^2.$$

Ostatecznie $P(S_n \neq S_n^*) \leqslant np^2 = \frac{\lambda^2}{n}$.

Wróćmy teraz do wypadków w armii pruskiej. Dzielimy rok na $n = 8760$ godzin i przyjmujemy, że szansa wypadku w korpusie w ciągu godziny jest równa $p = 0{,}000079908$, tak by $\lambda = np = 0{,}7$. Wygląda na to, że szansa dwóch lub większej liczby wypadków w ciągu godziny jest zaniedbywalnie mała. Zatem liczba wypadków ma z dobrym przybliżeniem rozkład Bernoulliego z parametrami n, p, a ten przybliża się rozkładem $Pois(\lambda)$ z dokładnością $\frac{\lambda^2}{n} = 0{,}000055936$.

Literatura

[1] Ladislaus von Bortkiewicz, *Das Gesetz der kleinen Zahlen*, B.G. Teubner, Leipzig 1898.

[2] Siméon D. Poisson, *Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédée des règles générales du calcul des probabilités*, 1837.