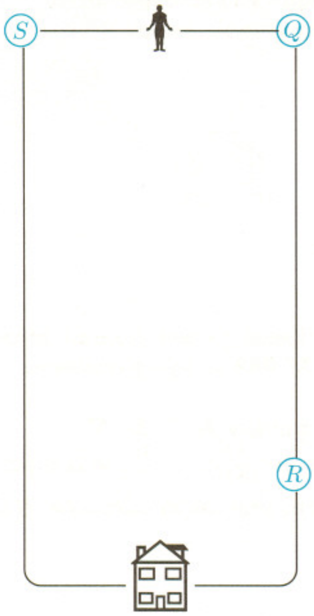
Średnia czy mediana?

Andrzej ZIELIŃSKI

Stałem kiedyś na rondzie między dwoma przystankami tramwajowymi i zastanawiałem się, który z nich wybrać (patrz rysunek 1). Do wyboru miałem przystanek S, z którego tramwaj bezpośrednio dowoził mnie do domu, lub przystanek Q, ale wtedy musiałbym przesiąść się na przystanku R. Ze względu na duże oddalenie przystanków S i Q musiałem zdecydować się na jeden z nich (stanie na środku i próba podbiegnięcia do jednego z nich w razie przyjazdu tramwaju nie wchodziła w rachubę).



Rozmyślania nad wyborem lepszego przystanku, czyli takiego, z którego szybciej dostanę się do domu, doprowadziły do powstania tego artykułu.

Oznaczmy przez x_S, y_Q, y_R czas czekania na przystanku S, Q i R odpowiednio (załóżmy, że są to zmienne losowe niezależne). Literami A, B i C oznaczmy odpowiednio: czas przejazdu z S do domu, czas przejazdu z Q do R i czas przejazdu z R do domu (załóżmy, że są to wielkości stałe).

Kryterium wyboru przystanku mogę sformułować następująco:

Przystanek S będzie lepszy od przystanku Q, gdy wybierając ten przystanek, stracę na dojazd średnio mniej czasu, niż gdybym wybrał przystanek Q, czyli gdy

$$E(x_S) + A \leq E(y_Q) + B + E(y_R) + C,$$

gdzie $E(x)$ to wartość średnia zmiennej x.

Wystarczy więc znać wszystkie wielkości występujące w powyższej nierówności, by problem był natychmiast rozstrzygnięty.

Niech $z = x_S - y_Q - y_R$, $D = B + C - A$ (z i D oznaczają odpowiednio różnice w czasie czekania i w czasie przejazdu między wariantami S i Q). Sformułowane kryterium można więc zapisać krótko:

$$E(z) \leq D.$$

Nazwijmy je *kryterium według wartości średniej.* Czy jest ono jednak dobre? Charakteryzuje się prostotą, ale nie uwzględnia – poza średnimi – rozkładów zmiennych x_S, y_Q i y_R.

Jeśli chcemy uwzględnić te rozkłady, kryterium wyboru przystanku moglibyśmy sformułować tak

Przystanek S będzie lepszy, gdy

$$P(x_S + A \leq y_Q + B + y_R + C) \geq P(x_S + A > y_Q + B + y_R + C),$$

czyli gdy $P(z \leq D) \geq \frac{1}{2}$.

To drugie kryterium związane jest z pewnym parametrem, który nazywa się *medianą.*

Definicja. Liczba m jest medianą zmiennej losowej x, jeśli $P(x \geq m) \geq \frac{1}{2}$ i $P(x \leq m) \geq \frac{1}{2}$.

Każda zmienna losowa ma medianę, ale nie zawsze jedną. Oznaczmy przez M_x zbiór wszystkich median zmiennej x. W dalszym ciągu rozpatrywać będziemy tylko te zmienne losowe, dla których M_x jest zbiorem jednopunktowym. (Jest tak np. dla rozkładu jednostajnego, normalnego, czy wykładniczego). W tym przypadku, dla podkreślenia jednoznaczności, jedyną medianę zmienej x oznaczać będziemy przez $M(x)$.

Drugie kryterium możemy teraz wyrazić następująco:

Przystanek S będzie lepszy, gdy $M(z) \leq D$.

Kryterium to nazwiemy kryterium względem mediany.

Powstaje pytanie, czy kryterium według średniej i kryterium względem mediany są równoważne? Mediana, w przeciwieństwie do średniej, na ogół nie jest addytywna, tzn. nie musi zachodzić wzór $M(x + y) = M(x) + M(y)$. Nawet więc

Rozwiązanie zadania F 554.
Wartość p_0 równa jest sumie ciśnienia powietrza w rurce nad rtęcią p_1 i ciśnienia hydrostatycznego słupa rtęci o wysokości H_1

$$p_0 = p_1 + H_1.$$

Podobnie, gdy wskazanie barometru jest równe H

$$p_{\text{atm}} = p + H.$$

Traktując powietrze w rurce jako gaz doskonały, można posłużyć się równaniem stanu

$$\frac{p_1(L - H_1)}{T_0} = \frac{p(L - H)}{T},$$

stąd

$$p = (p_0 - H_1)\frac{L - H_1}{L - H}\frac{T}{T_0}$$

i ostatecznie

$$p_{\text{atm}} = (p_0 - H_1)\frac{L - H_1}{L - H}\frac{T}{T_0} + H.$$

w najprostszym przypadku, gdy rozkłady x_S, y_Q i y_R są jednostajne, odpowiedź nie jest widoczna natychmiast. Bezpośrednie sprawdzenie daje odpowiedź pozytywną, co nie jest przypadkiem, zachodzi bowiem następujące

Twierdzenie. Jeśli zmienne losowe $x_1, x_2, \ldots, x_n$ są niezależne, każda ze zmiennych x_i, $i = 1, 2, \ldots, n$, ma rozkład symetryczny i jednoznaczną medianę $M(x_i)$, to również zmienna $x_1 + x_2 + \ldots + x_n$ ma jednoznaczną medianę oraz $M(x_1 + x_2 + \ldots + x_n) = M(x_1) + M(x_2) + \ldots + M(x_n)$.

Dowód twierdzenia. Wystarczy udowodnić twierdzenie dla rozkładów symetrycznych względem zera. Wtedy z założenia zmienna $u = x_1 + x_2 + \ldots + x_n$ ma rozkład symetryczny względem zera. Załóżmy, że jej mediana nie jest jednoznaczna, tzn. że medianą jest także pewna liczba $c > 0$. Ponieważ $P(u \le -c) = P(u \ge c) \ge 1/2$, więc $P(|u| < c) = 0$. Mamy

$$P\left(|x_1| < \frac{c}{n}\right) P\left(|x_2| < \frac{c}{n}\right) \ldots$$
$$\ldots P\left(|x_n| < \frac{c}{n}\right) =$$
$$= P\left(|x_1|<\frac{c}{n}, |x_2|<\frac{c}{n}, \ldots, |x_n|<\frac{c}{n}\right) \le$$
$$\le P(|u| < c) = 0.$$

Skąd wynika, że $P\left(|x_i| < \frac{c}{n}\right) = 0$ dla pewnego i, ale to oznacza, że $\frac{c}{n}$ jest różną od zera medianą zmiennej x_i. Sprzeczność z założeniem.

Przypomnijmy jeszcze, że zmienna x ma rozkład symetryczny względem zera, gdy zmienne x i $-x$ mają ten sam rozkład. Gdy taka zmienna ma wartość oczekiwaną, to wartość ta jest równa zeru. Gdy zmienne $x_1, x_2, \ldots, x_n$ mają rozkłady symetryczne względem zera, to również ich suma ma taki rozkład. Rozkład zmiennej x jest symetryczny, jeżeli istnieje takie a, że zmienna $x - a$ ma rozkład symetryczny względem zera. W takim przypadku, jeśli mediana jest jednoznaczna, to jest równa a.

Jasne jest teraz, że w naszym przypadku najciekawsza jest sytuacja, gdy zmienne x_S, y_Q i y_R nie są symetryczne. Przyjmijmy zatem, że zmienne te mają np. rozkład wykładniczy. Załóżmy też, że $E(y_Q) = E(y_R)$ oraz oznaczmy $k = \dfrac{E(y_Q)}{E(x_S)}$. Po żmudnych, niestety, rachunkach dojdziemy do następującego wyniku:

- jeśli $k \le \sqrt{2} - 1$, to $M(z) = \ln 2 - 2\ln(1 + k)E(x_S)$,
- jeśli $k > \sqrt{2} - 1$, to $M(z) = -sE(y_Q)$, gdzie s jest jedynym pierwiastkiem dodatnim równania $e^s = \dfrac{2k}{k+1}\left(s + \dfrac{k+2}{k+1}\right)$.

Powiedzmy, że czasy czekania na przystankach i czasy jazdy są następujące: $E(x_S) = 10$, $E(y_Q) = E(y_R) = 3$, $A = 22$, $B = 13$, $C = 11$ (wszystko, oczywiście, w minutach).

Wtedy $D = 2$, $E(z) = 4$, $M(z) = 1{,}68$.

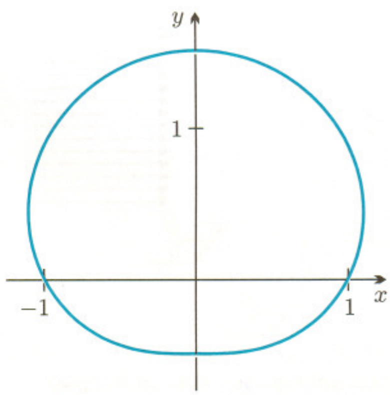
Przypomnijmy też, że rozkład wykładniczy (z parametrem $\lambda > 0$) to taki rozkład, dla którego zmienna przyjmuje tylko dodatnie wartości, przy czym prawdopodobieństwo tego, że zmienna osiągnie wartość z malutkiego przedziału długości $\mathrm{d}x$ zawierającego punkt $x > 0$, wynosi $\frac{1}{\lambda}e^{-x/\lambda}\mathrm{d}x$.

Jeśli czasy czekania na przystankach mają rozkłady wykładnicze, to który przystanek wybrać? Zgodnie z kryterium według wartości średniej powinienem wybrać przystanek Q, podczas gdy kryterium według mediany wskazuje na przystanek S. Postawię więc metapytanie: które kryterium jest lepsze? Otóż jeśli **częściej** chciałbym być w domu **szybciej**, to powinienem postąpić zgodnie z kryterium według mediany, czyli ustawiać się na przystanku S. Co wcale nie oznacza, że zyskałbym (np. w ciągu miesiąca takich przejazdów) na **łącznym czasie podróży**. Ten, zgodnie z kryterium według wartości średniej, byłby **krótszy** (średnio o 2 minuty), gdybym ustawiał się na przystanku Q.

Ale w takim razie wybór zależy np. od gustu, który – przynajmniej na razie i chyba na szczęście – nie poddaje się matematycznej formalizacji.

Droga Redakcjo

W marcowej tegorocznej *Delcie*, w artykule *Hipoteza, twierdzenie, dowód, kontrprzykład* znalazło się pytanie, czy znajomość długości cięciw przechodzących przez ustalony punkt wyznacza figurę wypukłą jednoznacznie. Odpowiedź na to pytanie jest negatywna.



Rozważmy koło o środku $O = (0,\ 0)$ i promieniu 1. Wszystkie cięciwy przechodzące przez O (czyli średnice) mają długość 2. Rozpatrzmy teraz figurę, której brzeg jest opisany we współrzędnych biegunowych przez równanie $r(\varphi) = (1 + \frac{1}{2}\sin\varphi)$ dla $\varphi \in \langle 0,\ 2\pi)$. Jest to figura wypukła, której kształt przedstawia rysunek.

Długość cięciwy o kierunku tworzącym z osią Ox kąt φ wynosi $(1 + \frac{1}{2}\sin\varphi) + (1 + \frac{1}{2}\sin(\pi + \varphi)) = (1 + \frac{1}{2}\sin\varphi) + (1 - \frac{1}{2}\sin\varphi) = 2$, niezależnie od φ. Mamy więc dwie nieprzystające figury o równych cięciwach.

Pozdrowienia

Witold BEDNAREK