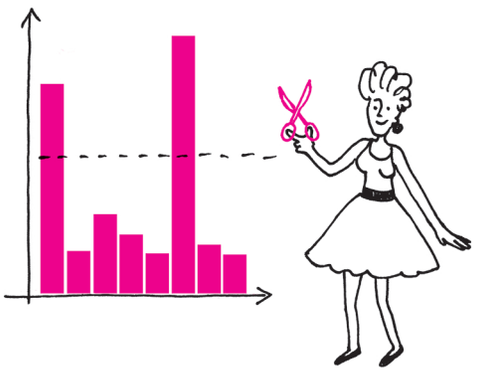
W krainie średnich

Przemysław GRZEGORZEWSKI[*]



Zdarza się czasami, że kiedy po przeprowadzeniu doświadczenia analizujemy dane, niektóre liczby wyglądają dziwnie – to znaczy inaczej, niż byśmy się spodziewali. W statystyce takie obserwacje, które są zdecydowanie większe lub zdecydowanie mniejsze od ogółu obserwacji nazywa się *obserwacjami odstającymi* (ang. *outliers*). Bardzo często pojawiają się one w próbie w wyniku błędów popełnionych w trakcie dokonywania obserwacji lub podczas późniejszego przetwarzania danych, na przykład przy wpisywaniu danych (przykładowo, wpisanie liczby 98 bądź 0,98 zamiast 9,8). Obecność takich błędnych wartości w próbce może wpływać niekorzystnie na wyniki obliczeń. W tym miejscu nasuwa się pytanie, czy nie należałoby usuwać ze zbioru danych „podejrzanych" wyników. Jeśli jesteśmy przekonani, że mamy faktycznie do czynienia z błędem, możemy to uczynić. Jednakże nie należy tego czynić pochopnie, bowiem obserwacje odstające nie zawsze są efektami błędów. Czasem są to wartości jak najbardziej prawidłowe, tyle że pojawiające się w danym doświadczeniu stosunkowo rzadko.

Mamy, na szczęście, do dyspozycji kilka prostych narzędzi statystycznych pozwalających poradzić sobie ze zniekształcaniem wyników pomiarów przez obserwacje odstające. Dla n-elementowego ciągu obserwacji $x_1, \ldots, x_n$ możemy obliczyć *średnią ważoną*, zdefiniowaną wzorem

$$\bar{x}_w = \sum_{i=1}^{n} w_i x_i, \tag{1}$$

gdzie ciąg nieujemnych wag w_i $(i = 1, \ldots, n)$ spełnia $w_1 + \ldots + w_n = 1$. Średnie ważone mogą różnić się nie tylko wartościami wag, ale i sposobem, w jaki wagi przypisywane są obserwacjom. W tradycyjnej średniej ważonej (1) i-ta waga w_i mnożona jest przez i-tą obserwację x_i, bez względu na to, jaką wartość przyjmuje x_i. W wielu sytuacjach wielce użyteczne bywają jednak takie średnie, dla których przydział wag jest uzależniony od uporządkowania obserwacji.

Uporządkowaną średnią ważoną, czyli w skrócie OWA (ang. *Ordered Weighted Average*), nazywamy średnią, w której i-ta waga w_i mnożona jest przez i-tą co do wielkości obserwację $x_{(i)}$, co możemy zapisać jako

$$\mathrm{OWA}(x, w) = \sum_{i=1}^{n} w_i x_{(i)}. \tag{2}$$

W statystyce wyrażenie (2) bywa nazywane L-statystyką. Szczególnym przypadkiem OWA jest tzw. *średnia ucięta*, której krańcowe pod względem wielkości obserwacje otrzymują wagę równą zeru. Średnią uciętą definiujemy więc jako

$$\bar{x}_{t,\alpha} = \frac{1}{n - 2\lfloor n\alpha \rfloor} \sum_{i=\lfloor n\alpha \rfloor + 1}^{n - \lfloor n\alpha \rfloor} w_i x_{(i)}, \tag{3}$$

gdzie $\alpha \geqslant 0$ jest tzw. wskaźnikiem ucięcia. Aby wzór (3) miał sens, wskaźnik ucięcia nie może być zbyt duży, a konkretnie, musi spełniać ograniczenie $\lfloor n\alpha \rfloor < n/2$. Indeks t występujący w symbolu średniej uciętej (2) pochodzi od angielskiej nazwy tej średniej, *trimmed mean*. To właśnie średnią uciętą stosuje się w praktyce po to, by wyeliminować wpływ ekstremalnych obserwacji.

W tym kontekście średnia ucięta może być postrzegana jako narzędzie pozwalające wyznaczyć przeciętną wartość badanej cechy w taki sposób, by nie tracąc z oczu obserwacji odstających, ignorować ich wpływ na dokonywane obliczenia. O średniej uciętej mówi się, iż jest ona odporna

[*]Wydział Matematyki i Nauk Informacyjnych, Politechnika Warszawska

Rozwiązanie zadania F 869.
Zadanie najwygodniej rozwiązywać w układzie związanym z obracającą się tarczą. Siła tarcia równa mgf, gdzie m jest masą pracownika, a g przyspieszeniem ziemskim, musi być przez cały czas ruchu nie mniejsza od wartości wypadkowej sumy sił: odśrodkowej i Coriolisa. Siła odśrodkowa i siła Coriolisa są do siebie prostopadłe i co do wartości odpowiednio równe $m\omega^2 r$ oraz $2\omega v$. Prędkość pracownika (zgodna z warunkami zadania) to $v = R\omega/\pi$.

Stąd warunek powodzenia zamiaru pracownika to:

$$g^2 f^2 \geqslant \omega^4 \left(r^2 + 4\frac{R^2}{\pi^2} \right)$$

dla każdej odległości r pracownika od środka tarczy. Ostatecznie:

$$g^2 f^2 \geqslant \omega^4 R^2 \left(1 + \frac{4}{\pi^2} \right).$$

na obecność obserwacji odstających w próbce. Tej własności nie ma natomiast zwykła średnia arytmetyczna, która przykłada taką samą wagę do wszystkich obserwacji, w tym także odstających.

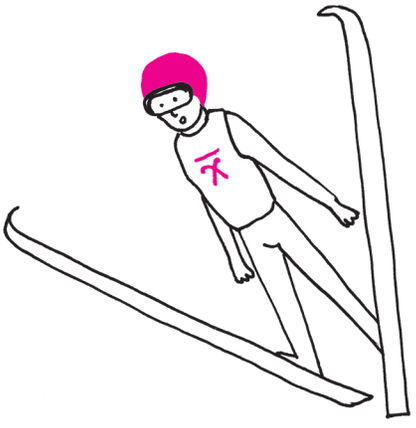
Ilustracją stosowania średniej uciętej w życiu codziennym może być sposób oceny skoczków narciarskich. Jak wiadomo, skok oceniany jest przez pięciu sędziów, przy czym faktyczna ocena skoku dokonywana jest na podstawie trzech ocen pozostałych po wyeliminowaniu dwóch ocen ekstremalnych – minimalnej i maksymalnej. Celem takiego postępowania jest uniknięcie sytuacji, w której sędzia miałby faworyzować skoczka wystawiając mu notę dużo wyższą niż pozostali sędziowie, bądź też chciałby ocenić skok dużo gorzej niż inni jurorzy.

Dopuszczając maksymalny stopień ucięcia otrzymujemy średnią uciętą postaci

$$\text{med} = \begin{cases} x_{((n+1)/2)} & \text{dla } n \text{ nieparzystego,} \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right) & \text{dla } n \text{ parzystego.} \end{cases} \tag{4}$$

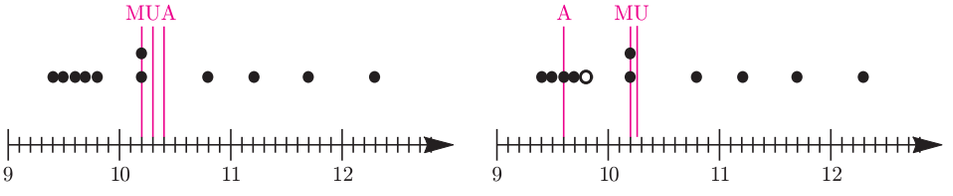
Ten szczególny przypadek średniej uciętej nazywamy *medianą*.

Mediana w sposób oczywisty jest odporna na wartość obserwacji odstających, bowiem, jak widać ze wzoru (4), bazuje wyłącznie na obserwacjach położonych w samym środku uporządkowanej niemalejąco próbki. W pewnych sytuacjach bywa to zaletą, ale trzeba pamiętać, że stosując zbyt duże ucięcie pozbywamy się nie tylko obserwacji odstających, ale i wielu cennych informacji zawartych w próbce. Tak więc w przypadku średniej uciętej, podobnie jak w życiu codziennym, trzeba postępować rozważnie i nie popadać w skrajności.



Jeszcze innym rodzajem średniej, używanej w statystyce i odpornej na obecność obserwacji odstających, jest tzw. *średnia winsorowska*. Wyznacza się ją w sposób podobny do średniej uciętej, tyle że zamiast eliminowania krańcowych obserwacji z uporządkowanego ciągu, zastępuje się je odpowiednio dobranymi wartościami. Dokładniej, po uporządkowaniu zbioru obserwacji w sposób niemalejący, ucinamy ustaloną liczbę krańcowych obserwacji (od dołu i od góry), a następnie dopisujemy tyle obserwacji, ile łącznie ucięliśmy, przy czym połowa spośród dopisanych obserwacji ma być równa pierwszej, a połowa ostatniej obserwacji z podzbioru, który pozostał po ucięciu. W przypadku średniej winsorowskiej, w przeciwieństwie do średniej uciętej, mamy do czynienia nie tyle z eliminacją obserwacji odstających, co raczej z ograniczeniem ich wpływu na wyznaczaną wartość przeciętną.

Jak widać, pojęcie wartości przeciętnej, czy też średniej, mimo iż tak intuicyjne, niekoniecznie musi prowadzić do średniej arytmetycznej. Jakkolwiek jest ona powszechnie i często z powodzeniem stosowana w praktyce, zdarzają się i takie sytuacje, w których jej użycie może przynieść niepożądane efekty. Pamiętajmy, że w wielu sytuacjach dobór właściwej średniej może się okazać decydujący dla poprawności uzyskiwanych wniosków.



Przykładowe 11 wartości wyników pomiarów oraz ich średnia arytmetyczna (A), mediana (M) i średnia ucięta z parametrem ucięcia $\alpha = 0{,}1$ (U). Zestaw danych na rysunku po prawej stronie różni się od tego po lewej tym, że wartość pomiaru zaznaczonego pustym kółkiem została (przypadkowo?) zmniejszona dziesięciokrotnie.