Andrzej Sierociński

STATYSTYKA MATEMATYCZNA

Elementy Statystyki Opisowej

Wykład 10

27 Elementy statystyki opisowej

27.1 Tablica częstości (szereg rozdzielczy), histogram, łamana częstości

Niech $x_1, x_2, \ldots, x_n$ będzie n-elementową próbką.

Rozstęp
Rozstępem z próbki nazywamy

$$R = \max\{x_1, x_2, \ldots, x_n\} - \min\{x_1, x_2, \ldots, x_n\}$$

Przy większej liczności próbki ($n > 30$), w celu ułatwienia analizy danych, wartości liczbowe próbki grupuje się w klasach (najczęściej o jednakowej długości), przyjmując uproszczone założenie, że wszystkie wartości znajdujące się w danej klasie są identyczne ze środkiem klasy.

Liczba klas - k.

Liczność próbki n	Liczba klas k
30 - 60	6 - 8
60 - 100	7 -10
100 - 200	9 -12
200 - 500	11 - 17
500 - 1500	16 - 25

Na ogół nie stosuje się liczby klas k większej od 30.

Długość klasy - b.

$$b \cong \frac{R}{k} \quad (b \cdot k \geqslant R)$$

Punkty stanowiące granice poszczególnych klas (zwykle) ustala się z dokładnością do $\alpha/2$, gdzie α jest dokładnością pomiaru.

Oznaczmy przez n_i liczność i-tej klasy. Oczywiście $\sum_{i=1}^{k} n_i = n$.

Szereg rozdzielczy
Szeregiem rozdzielczym nazywamy ciąg par $(\bar{x}_i, n_i)$, $i = 1, \ldots, k$, gdzie $\bar{x}_i$ jest środkiem i-tej klasy.

Ciąg $\{n_i\}$ nazywamy rozkładem liczności badanej cechy przy danej liczbie k klas.

Przykład *Wkładka topikowa bezpiecznika o natężeniu znamionowym 20A winna, zgodnie z normą, wytrzymać bez przepalenia się natężenie 28A w ciągu 1 godziny. W celu sprawdzenia zgodności z normą, z partii wkładek topikowych tego typu pobrano losowo 40 sztuk i zanotowano czasy przepalenia się wkładki przy natężeniu prądu 28A. Otrzymano następujące wyniki w minutach:*

51	*58*	*64*	*69*	*61*	*56*	*41*	*48*	*56*	*61*
75	*55*	*46*	*57*	*70*	*55*	*47*	*62*	*55*	*60*
54	*57*	*65*	*60*	*53*	*54*	*49*	*58*	*62*	*59*
53	*50*	*58*	*63*	*64*	*59*	*52*	*51*	*65*	*60*

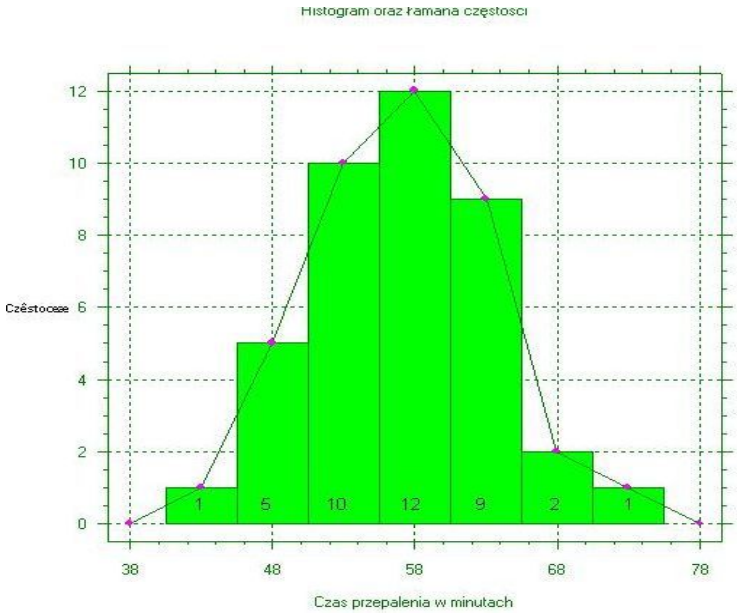
Dla przedstawionej próbki zbudować szereg rozdzielczy oraz narysować histogram i łamaną częstości.

Rozwiązanie *Zauważmy, że* $x_{min} = 41$ *oraz* $x_{max} = 75$ *. Zatem rozstęp z próbki* $R = 34$*. Ponieważ liczność próbki* $n = 40$*, to wygodnie jest przyjąć liczbę klas* $k = 7$ *oraz szerokość klasy* $b = 5$*.*

Tym samym otrzymujemy następujący szereg rozdzielczy:

Nr klasy i	*Klasa*	$\bar{x}_i$	n_i	w_i	N_i	W_i
1	*40.5 - 45.5*	*43.0*	*1*	*.0250*	*1*	*.0250*
2	*45.5 - 50.5*	*48.0*	*5*	*.1250*	*6*	*.1500*
3	*50.5 - 55.5*	*53.0*	*10*	*.2500*	*16*	*.4000*
4	*55.5 - 60.5*	*58.0*	*12*	*.3000*	*28*	*.7000*
5	*60.5 - 65.5*	*63.0*	*9*	*.2250*	*37*	*.9250*
6	*65.5 - 70.5*	*68.0*	*2*	*.0500*	*39*	*.9750*
7	*70.5 - 75.5*	*73.0*	*1*	*.0250*	*40*	*1.0000*

gdzie $N_i = \sum_{i=1}^{k} n_i$ *oraz* $W_i = \sum_{i=1}^{k} w_i$ *są licznościami i częstościami łącznymi odpowiednio.*



27.2 Miary opisowe (wyznaczanie wielkości najbardziej reprezentatywnych rozproszenia, skośności i spłaszczenia badanej cechy)

Problem promocji. *Załóżmy, że stajemy przed następującym problemem promocji. W wyniku odejścia jednego z kierowników zwolniło się odpowiednie stanowisko i na to miejsce chcemy przeszeregować jednego z naszych pracowników. Wszyscy pracownicy naszej firmy po pierwszym roku pracy przechodzą test mający ocenić ich przydatność na stanowisku kierowniczym. Jednakże ostatnio sam test uległ zmianie i część pracowników jest oceniona według nowej skali ocen. Zasadą jest, że promuje się pracownika, który osiągnął najlepszy wynik.*
Okazało się, że jest dwóch kandydatów o wynikach 143 (nowy test) oraz 29 (stary test). Oczywiście nie byłoby w porządku poddawanie nowemu testowi pracownika, który osiągnął wynik 29.

1. *W jaki sposób porównać te dwa wyniki?*
2. *Jakie dodatkowe informacje musimy uzyskać, żeby to porównanie było możliwe?*

27.2.1 Miary położenia (miary środka rozkładu - wyznaczanie wielkości najbardziej reprezentatywnych)

Średnia (Average)

$$\begin{aligned} \bar{x}_n &= \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{dla danych “surowych” (niepogupowanych)} \\ \bar{x}_n &= \frac{1}{n}\sum_{j=1}^{k} n_j \bar{x}_j \quad \text{dla danych pogrupowanych w } k \text{ klas} \end{aligned}$$

Średnią nazywamy **miarą klasyczną**. Nietrudno zauważyć, że jej wartość jest wartością oczekiwaną zmiennej losowej o dyskretnym rozkładzie jednostajnym na zbiorze wartości próby.

Inną miarą środka rozkładu jest mediana, która jest tzw. **miarą pozycyjną**.

Mediana

$$\begin{aligned} x_{med} &= \frac{x_{(\lceil \frac{n}{2} \rceil)} + x_{(\lceil \frac{n+1}{2} \rceil)}}{2} \quad \text{dla danych niepogupowanych} \\ x_{med} &= x_l + \frac{b}{n_m}\left(\frac{n}{2} - \sum_{j=1}^{m-1} n_j\right) \quad \text{dla danych pogrupowanych} \end{aligned}$$

gdzie m oznacza numer klasy mediany, x_l lewy koniec klasy mediany a $x_{min} = x_{(1)} \leqslant x_{(2)} \leqslant x_{(3)} \leqslant \cdots \leqslant x_{(n)} = x_{max}$ uporządkowaną (posortowaną) próbkę.

Moda Wartością modalną (modą, dominantą) x_{mod} próbki $x_1, x_2, \ldots, x_n$ o powtarzających się wartościach nazywamy najczęściej powtarzającą się wartość, o ile istnieje, nie będącą x_{min} ani też x_{max}.

W przypadku danych pogrupowanych modą nazywamy środek najliczniejszej klasy za wyjątkiem klas skrajnych.

Jeżeli w szeregu rozdzielczym najliczniejszymi są obie klasy skrajne, to szereg rozdzielczy nazywamy antymodalnym **typu U**, a środek najmniej licznej klasy **antymodą**.

Gdy najliczniejsza jest jedna z klas skrajnych, wtedy szereg rozdzielczy nazywamy **antymodalnym typu J**. W przypadku, gdy istnieje więcej niż jedna wartość modalna to rozkład takiej cechy nazywamy rozkładem **wielomodalnym**.

Przykład (cd.) *Wyznaczymy parametry położenia najpierw dla danych niepogupowanych.*

$$\bar{x} = \frac{1}{40}\sum_{i=1}^{40} x_i = 57,325.$$

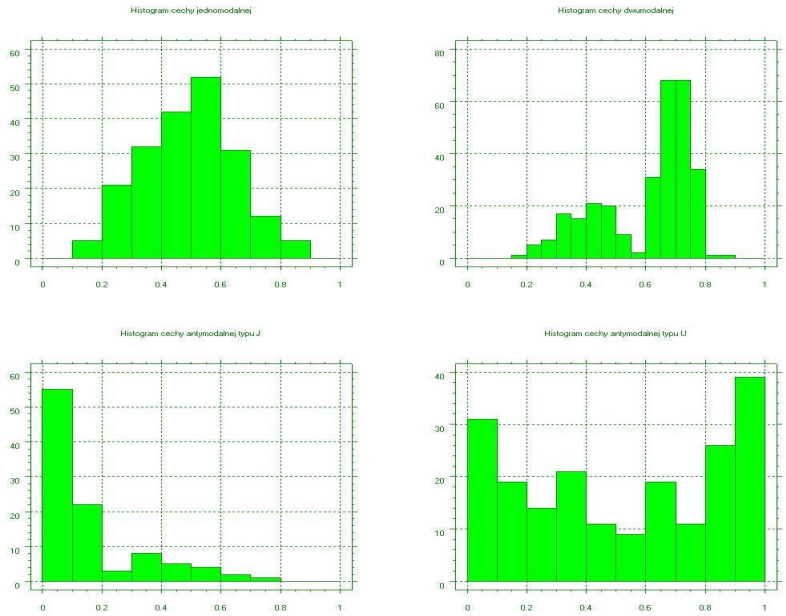
$$x_{med} = \frac{x_{(\lceil \frac{40}{2} \rceil)} + x_{(\lceil \frac{41+1}{2} \rceil)}}{2} = \frac{x_{(\lceil 20 \rceil)} + x_{(\lceil 21 \rceil)}}{2} = \frac{57+58}{2} = 57,5.$$

Dla danych pogrupowanych mamy

$$\bar{x} = \frac{1}{40} \sum_{j=1}^{7} n_j \bar{x}_j = 57,125$$

$$x_{med} = 55,5 + \frac{5}{12} \left(\frac{40}{2} - \sum_{j=1}^{4-1} n_j \right) = 57,167.$$

Moda w obu przypadkach jest taka sama $x_{mod} = 58$.

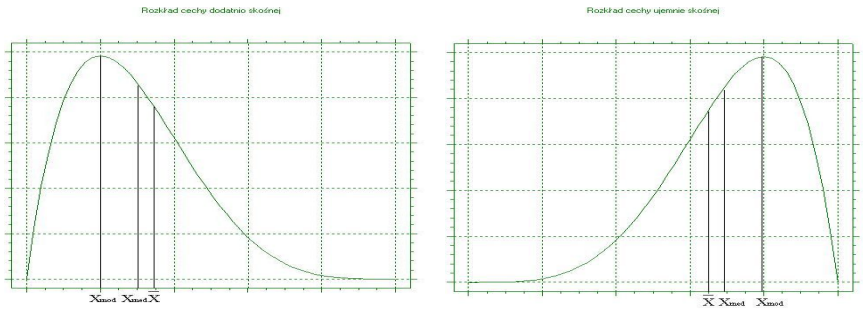


Dla próbki z populacji o **rozkładzie symetrycznym** wszystkie miary środka rozkładu dają wartości zbliżone, tzn.

$$\bar{x}_n \cong x_{med} \cong x_{mod}.$$

Natomiast dla **rozkładów asymetrycznych** mamy następującą zależność empiryczną:

$$|\bar{x}_n - x_{mod}| \cong 3 \cdot |\bar{x}_n - x_{med}|$$



27.2.2 Miary rozproszenia

Wariancja Wariancją z próby nazywamy

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \quad \text{dla danych niepogupowanych}$$

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{k} n_j\,(\bar{x}_j - \bar{x}_n)^2 \quad \text{dla danych pogrupowanych.}$$

Odchylenie standardowe Odchyleniem standardowym z próby nazywamy

$$s = \sqrt{s^2}.$$

Zarówno wariancja jak i odchylenie standardowe są miarami klasycznymi. Do najczęściej używanych miar pozycyjnych należą **rozstęp, rozstęp międzykwartylowy i odchylenie pseudostandardowe**.

Rozstęp z próby Rozstępem z próby nazywamy

$$R = x_{max} - x_{min} = x_{(n)} - x_{(1)}.$$

Kwartyle Kwartylami dolnym i górnym z próby nazywamy odpowiednio

$$Q_1 = \frac{x_{(\lceil \frac{n}{4} \rceil)} + x_{(\lceil \frac{n+1}{4} \rceil)}}{2}, \qquad Q_3 = \frac{x_{(\lceil \frac{3n}{4} \rceil)} + x_{(\lceil \frac{3n+1}{4} \rceil)}}{2}.$$

Rozstęp międzykwartylowy z próby Rozstępem międzykwartylowym z próby nazywamy

$$IQR = Q_3 - Q_1.$$

W podobny sposób można zdefiniować dowolne kwantyle z próby.

p-ty q-ty kwantyl p-tym q-tym kwantylem z próby nazywamy

$$q_{p,q} = \frac{x_{(\lceil \frac{pn}{q} \rceil)} + x_{(\lceil \frac{pn+1}{q} \rceil)}}{2}.$$

Na przykład, dla $p = 1$ i $q = 2$ mamy medianę, a dla $p = 1$ i $p = 3$ oraz $q = 4$ oba kwartyle.

Dla $q = 10$ mamy **decyle** a dla $q = 100$ **percentyle** z próby.

Jeżeli zmienna losowa ma dowolny rozkład normalny $N(\mu, \sigma^2)$, to teoretyczny rozstęp międzykwartylowy wynosi $IQR = 1,35 \cdot \sigma$ niezależnie od wartości oczekiwanej μ. Dlatego częstokroć do oceny odchylenia standardowego z próby przyjmuje się

Odchylenie pseudostandardowe z próby Odchyleniem pseudostandardowym z próby nazywamy

$$PSD = \frac{IQR}{1,35}$$

Jeżeli $PSD < s$ to badana cecha ma "tłuste ogony". W przypadku $PSD > s$ rozkład ma "wybrzuszenie" w środku. Tylko w przypadku, gdy $PSD \cong s$ oraz rozkład częstości jest symetryczny, można uznać, że dane pochodzą z rozkładu normalnego.

Przykład (cd.) *Dla danych dotyczących czasów przepalenia się wkładki topikowej otrzymujemy, że*

$$s^2 = 47,353 \quad oraz \quad s = 6,881$$

w przypadku danych niepogupowanych oraz

$$s^2 = 40,881 \quad oraz \quad s = 6,394$$

dla danych pogrupowanych. Jak łatwo zauważyć, grupowanie danych spowodowało znaczne niedoszacowanie rzeczywistej wariancji.

Z histogramu oraz małych różnic pomiędzy średnią, medianą i modą wynika, że rozkład badanej cechy jest symetryczny. Wyznaczmy odchylenie pseudostandardowe. Nietrudno sprawdzić, że

$$Q_1 = 53 \quad oraz \quad Q_3 = 61,5.$$

Zatem

$$PSD = \frac{61,5 - 53}{1,35} = 6,296.$$

Potwierdza to nasze przypuszczenie o symetrii rozkładu.

Jeżeli chcemy porównać dwie cechy losowe, które są wyrażone w innych jednostkach lub różnią się między sobą rzędem wielkości, to musimy zrezygnować z porównania wariancji lub odchylenia standardowego, ponieważ ulegają one zmianie wraz ze zmianą skali.

W takim przypadku używamy współczynnika, który mierzy rozproszenie względne.

Współczynnik zmienności z próby Współczynnikiem zmienności z próby nazywamy

$$CV = \frac{s}{\bar{x}} \cdot 100\%$$

Współczynnik zmienności pozwala na

1. pomiar względnej zmienności (na jednostkę średniej);
2. umożliwia porównanie zmienności dwu lub więcej zbiorów danych, także wyrażonych w różnych jednostkach.

Przykład. *W pewnej korporacji odchylenie standardowe zarobków w grupie pracowników niższego szczebla oraz w grupie menadżerów jest jednakowe i wynosi 500zł. Czy można twierdzić, że zróżnicowanie zarobków w obu grupach pracowników jest jednakowe?*

Rozwiązanie
Aby prawidłowo odpowiedzieć na to pytanie należy znać średnie zarobki w obu grupach pracowników. Wiadomo, że średnie wynagrodzenie w grupie pracowników niższego szczebla (A) wynosi 3500zł, natomiast w grupie menadżerów (B) 12000zł.

Zatem współczynnik zmienności wynosi

$$CV_A = \frac{500}{3500} * 100\% = 14,29\%$$

dla grupy pracowników niższego szczebla, oraz

$$CV_B = \frac{500}{12000} * 100\% = 4,17\%.$$

Wynika stąd, że w grupie menadżerów mamy znacznie mniejsze zróżnicowanie (względne) zarobków.

27.2.3 Standaryzacja wartości w próbie (z-score)

W przypadku, gdy chcemy ze sobą porównać dwie wartości cechy mierzone w różnych skalach, musimy najpierw dokonać standaryzacji, polegającej na wyrażeniu tych wartości w wielkościach sprowadzonych do tej samej skali.

Standaryzacja wartości w próbie (z-score) Wartością standaryzowaną wielkości $x_i, i = 1, 2, \ldots, n$ w próbie $x_1, x_2, \ldots, x_n$, nazywamy

$$z_i = \frac{x_i - \bar{x}_n}{s}.$$

Operacja ta jest operacją przeskalowania danych. W przypadku dwóch różnych próbek o rozkładach symetrycznych zbliżonych do rozkładu normalnego pozwala na porównywanie między sobą ich wartości.

Rozwiązanie problemu promocji.
Z zebranych wyników pierwszego testu mamy $\bar{x}' = 22.6$, $x'_{med} = 22.9$ *oraz* $s' = 2.8$. *Podobnie osoby, które przeszły drugi test uzyskały wyniki* $\bar{x}'' = 107.8$, $x''_{med} = 104.9$ *oraz* $s'' = 17.4$.

W obu przypadkach testowi poddano po kilkuset pracowników.

Ponieważ zarówno w pierwszym jak i w drugim przypadku można przyjąć, że oceny mają rozkłady symetryczne, obliczmy wartości standaryzowane dla obu wybranych pracowników:

$$z' = \frac{29 - 22,6}{2,8} = 2,29 \quad \textit{oraz} \quad z'' = \frac{143 - 107,8}{17,4} = 2,02.$$

Z porównania obu liczb wynika, że pierwszy pracownik uzyskał relatywnie lepszą ocenę.

27.2.4 Dystrybuanta empiryczna, momenty empiryczne

Nietrudno zauważyć, że klasyczne miary położenia oraz rozproszenia odpowiadają wartościom momentów pierwszego i drugiego rzędu, dyskretnego rozkładu jednostajnego na zbiorze punktów próby $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$.

Nic nie stoi na przeszkodzie, żeby zdefiniować w podobny sposób kurtozę i współczynnik asymetrii z próby.

Najpierw zdefiniujemy pojęcie **dystrybuanty empirycznej** jako dystrybuanty rozkładu jednostajnego na $\mathcal{X}$.

Dystrybuanta empiryczna
Dystrybuantą empiryczną nazywamy funkcję $F_n : \mathcal{R} \to [0, 1]$

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}\{x_i \leqslant x\},$$

gdzie $\mathbf{I}\{x_i \leqslant x\}$ jest funkcją wskaźnikową zbioru $(-\infty, x_i]$, a $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ próbą losową.

Empiryczny moment zwykły rzędu l

$$\begin{aligned} m_l &= \frac{1}{n} \sum_{i=1}^{n} x_i^l \quad \text{dla danych niepogupowanych} \\ m_l &= \frac{1}{n} \sum_{j=1}^{k} n_j \bar{x}_j^l \quad \text{dla danych pogrupowanych w } k \text{ klas} \end{aligned}$$

Empiryczny moment centralny rzędu l

$$\begin{aligned} M_l &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^l \quad \text{dla danych niepogupowanych} \\ M_l &= \frac{1}{n} \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x}_n)^l \quad \text{dla danych pogrupowanych w } k \text{ klas} \end{aligned}$$

Nietrudno zauważyć, że średnia $\bar{x} = m_1$ oraz wariancja $s^2 = \frac{n}{n-1} M_2$.

Empiryczny współczynnik skośności (skewness)
Empirycznym współczynnikiem skośności nazywamy

$$g_1 = \frac{M_3}{s^3}.$$

Współczynnik ten charakteryzuje skośność rozkładu badanej cechy. W praktyce przyjmuje się, że dla $|g_1| < 0,5$ rozkład jest symetryczny, natomiast dla $|g_1| > 1$ mocno skośny. Znak współczynnika wskazuje, w którą stronę jest skośna cecha. Dla wartości dodatniej mamy skośność w prawo a dla ujemnej w lewo.

W przypadku, gdy chcemy uniknąć dość kłopotliwego wyznaczania trzeciego momentu centralnego, do badania skośności można użyć **indeksu skośności Pearsona**. Indeks ten jest wygodną oceną skośności w sytuacji, gdy nie dysponujemy specjalnym programem obliczeniowym. Wartości indeksu I interpretuje się podobnie jak wartości współczynnika g_1.

Indeks skośności Pearsona

$$I = \frac{3(\bar{x} - x_{med})}{s}$$

lub

$$I = \frac{\bar{x} - x_{mod}}{s}.$$

Empiryczny współczynnik spłaszczenia - kurtoza
Empiryczną kurtozą nazywamy

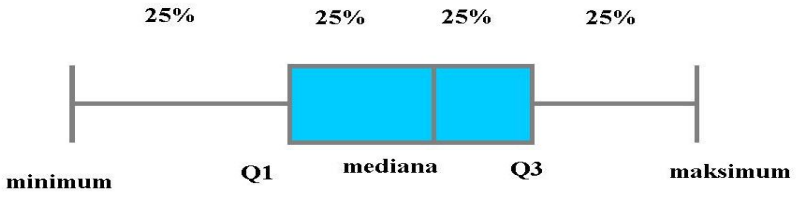
$$g_2 = \frac{M_4}{s^4} - 3.$$

Dla cechy o rozkładzie normalnym lub podobnym kurtoza w przybliżeniu jest równa zeru. Wartości mniejsze od zera świadczą o spłaszczeniu rozkładu w porównaniu do rozkładu normalnego, natomiast wartości dodatnie o tym, że rozkład ma pik.

W przypadku małych prób, o liczności $n < 30$, grupowanie w klasy nie ma większego sensu.

Trudno oczekiwać, aby histogram mógł oddawać w pełni charakter rozkładu badanej cechy. Dlatego John Tukey wprowadził tzw. wykres "pudełkowy" (Box and whiskers plot), który oparty jest na pięciu wskaźnikach sumarycznych (five numbers summary): minimum, maksimum,dolny i górny kwartyl oraz mediana.

Na poniższym rysunku przedstawimy jego wersję uproszczoną.



27.2.5 Wykrywanie wartości nietypowych w próbie (outliers)

Większość standardowych metod wnioskowania statystycznego zakłada, że mamy do czynienia z próbką z rozkładu normalnego. Ponieważ praktycznie 100% obserwacji z populacji o rozkładzie normalnym zawiera się w przedziale $[-3s, +3s]$, to obserwacje nie wpadające do tego przedziału traktowane są jako obserwacje nietypowe.

Jeśli nie jest prawdą, iż cecha ma rozkład normalny to obliczone z próbki wartości średniej i odchylenia standardowego nie dają pełnego obrazu rozkładu badanej cechy.

Zaobserwowanie nietypowych wartości ekstremalnych w próbce, tzw. **outliers** może spowodować problem w ich interpretacji. Otóż wartości nietypowe mogą sygnalizować fakt, iż próbka nie pochodzi z rozkładu normalnego lub, że nastąpił błąd przy zbieraniu danych (zły pomiar lub błąd w zapisie).

W większości przypadków wartości nietypowe są wynikiem rzeczywistego mechanizmu losowego i nie można ich pomijać z rozważań. Jednak wartości średniej i odchylenia standardowego mogą być obciążone błędem.

Jeżeli wartości nietypowe skupiają się po jednej stronie średniej to następuje przesunięcie średniej, jeżeli są rozłożone symetrycznie to średnia może być dobrym oszacowaniem środka rozkładu, ale odchylenie standardowe może być zbyt duże.

Zarówno średnia jak i odchylenie standardowe nie są odporne na efekt występowania wartości nietypowych.

Pierwszym krokiem przy sprawdzaniu normalności badanej cechy jest ustalenie czy wśród danych zebranych w próbce występują wartości nietypowe, jeśli tak to czy można je przypisać popełnionym błędom w trakcie zbierania danych.

Jeśli nie, to znaczy że cecha ma rozkład skośny (wartości nietypowe układają się po jednej stronie) lub ma "długie ogony" (long-tailed). W obu przypadkach używanie średniej i odchylenia standardowego do oceny odpowiednich parametrów jest bardzo ryzykowne.

Najprostszym sposobem wykrywania wartości nietypowych jest stwierdzenie czy leżą w przedziale trzech odchyleń standardowych wokół średniej, tzn. czy wartości po standaryzacji są większe co do wartości bezwzględnej od 3. Takie wartości nazywamy **ekstremalnie nietypowymi**, ponieważ szansa zaobserwowania jest bliska zeru (dla zmiennej losowej o rozkładzie normalnym wynosi 0,0027).

Większość pakietów statystycznych jako nietypowe określa wartości dla których standaryzowana wartość $|z_i| > 2$, ponieważ dla obserwacji normalnych szansa zaobserwowania jest mniejsza od 5%.

Jednak jak to zostało stwierdzone powyżej takie postępowanie może być obarczone błędem.

Inne podejście do tego problemu zaproponował Tukey jest to tzw. Box-and-Whisker Plot, który został omówiony wcześniej. W tej metodzie przyjmuje się zasadę, że obserwacja, dla której

$$x_i \leqslant Q_1 - 1,5 \cdot IQR \quad \text{lub} \quad x_i \geqslant Q_3 + 1,5 \cdot IQR$$

uznawana jest za nietypową, a jeżeli

$$x_i \leqslant Q_1 - 3 \cdot IQR \quad \text{lub} \quad x_i \geqslant Q_3 + 3 \cdot IQR,$$

to za ekstremalnie nietypową.