

Statystyka opisowa

Statystyka opisowa różni się od statystyki matematycznej tym, że próby statystycznej dotyczącej danej cechy, nie wykorzystuje się do wnioskowania na temat populacji, z której próba ta została wylosowana, a jedynie analizuje się zależności w próbie. W statystyce opisowej nie używa się więc narzędzi wykształconych przez rachunek prawdopodobieństwa – używa się natomiast niezbyt skomplikowanych technik służących poznaniu próbkowych, ogólnych własności rozkładu danej cechy.

W szczególności, przedmiotem zainteresowania statystyki opisowej są przede wszystkim:

1. miary położenia (m.in. tendencji centralnej): np. średnia, dominanta (moda), kwantyle.
2. miary dyspersji (rozproszenia, zróżnicowania): np. wariancja, odchylenie standardowe, odchylenie przeciętne.
3. miary asymetrii (skośności): np. klasyczny/pozycyjny współczynnik asymetrii.

Wspomniane miary dzielimy na *klasyczne* i *pozycyjne*. Miary klasyczne, to takie, które wykorzystują wszystkie wartości cechy w próbie. Przykładem takich miar są: momenty zwykłe (m.in. średnia arytmetyczna), wariancja, klasyczny współczynnik zmienności.

Miary pozycyjne wykorzystują wartości tylko niektórych jednostek z próby i są alternatywą dla miar klasycznych. Przykładem miar pozycyjnych jest dominanta i kwantyle.

Dane statystyczne mogą być zaprezentowane w różny sposób – od tego sposobu zależeć też będzie rodzaj ich analizy. Dajmy na to, że mamy wyniki badania wzrostu w pewnej drużynie koszykówki:

188, 175, 199, 209, 188, 194, 195, 195, 202, 191, 203, 188, 203, 193, 197, 198, 185, 195, 194, 202.

Powyższy sposób prezentacji danych, który pokazuje wartość każdej obserwacji w próbie z osobna, to tzw. *szereg szczegółowy (indywidualny)*.

Te same dane można również zaprezentować w *szeregu rozdzielczym punktowym*:

x_i	n_i
175	1
185	1
188	3
191	1
193	1
194	2
195	3
197	1
198	1
199	1
202	2
203	2
209	1

W szeregu tym prezentowana jest ilość wystąpień w próbie każdej wartości cechy.

W końcu, te same dane można przedstawić w *szeregu rozdzielczym przedziałowym*. Ten sposób prezentacji danych pokazuje ile jest wystąpień wartości cechy należących do danego *przedziału klasowego*.

Przykładowo, w naszym przypadku, przyjmując równej długości przedziały klasowe, dane można by zaprezentować następująco:

x_{0i}	x_{1i}	n_i
[175	185)	1
[185	195)	8
[195	205)	10
[205	215)	1

Istnieją różne reguły ustalania liczby przedziałów klasowych. Najczęściej wykorzystywanymi są:

1. $k \approx \sqrt{n}$
2. $k \approx 1 + 3,32 \ln(n)$
3. $k \leq 5 \ln(n)$

W praktyce jednak reguły te nie są zbyt często stosowane i ostateczna liczba przedziałów najczęściej zależy od zdroworozsądkowego podejścia statystyka do konkretnego zagadnienia.

Wzory:

0. Momenty (wzory zaprezentowane dla szeregów rozdzielczych punktowych – łatwo je zmodyfikować, żeby obejmowały szeregi szczegółowe lub rozdzielcze przedziałowe).

a. momenty zwykłe

Wzór ogólny (moment zwykły rzędu r):

$$m_r(x) = \frac{\sum_{i=1}^p x_i^r n_i}{n}, \text{ gdzie } p \text{ to ilość różnych wartości analizowanej cechy.}$$

Przykładowo $m_1(x) = \frac{\sum_{i=1}^p x_i n_i}{n}$ to po prostu średnia arytmetyczna.

b. momenty centralne

Wzór ogólny (moment centralny rzędu r):

$$\mu_r(x) = \frac{\sum_{i=1}^p (x_i - \bar{x})^r n_i}{n}$$

Przykładowo trzeci moment centralny $\mu_3(x) = \frac{\sum_{i=1}^p (x_i - \bar{x})^3 n_i}{n}$ służy do oceny asymetrii rozkładu cechy.

1. Miary położenia

a. średnia arytmetyczna:

i. szereg szczegółowy: $\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$

ii. szereg rozdzielczy punktowy: $\bar{X} = \frac{\sum_{i=1}^p x_i n_i}{n}$, gdzie p oznacza ilość różnych wartości cechy.

iii. szereg rozdzielczy przedziałowy: $\bar{X} = \frac{\sum_{i=1}^k \overset{o}{x_i} n_i}{n}$, gdzie k oznacza ilość przedziałów klasowych, zaś $\overset{o}{x_i}$ jest środkiem i -tego przedziału.

b. dominanta (moda) – najczęściej występująca wartość

i. szereg rozdzielczy punktowy: $D(x) = x_{i \ (i: n_i = \max_i(n_i))}$

ii. szereg rozdzielczy przedziałowy:

$$D(x) = x_{0d} + \frac{(n_d - n_{d-1})}{(n_d - n_{d-1}) + (n_d - n_{d+1})} \cdot h_d, \text{ gdzie } n_d, n_{d-1} \text{ i } n_{d+1} \text{ to}$$

odpowiednio liczebności przedziału dominanty, przedziału poprzedzającego dominantę i przedziału następującego po dominancie, h_d to rozpiętość przedziału dominanty, zaś x_{0d} to dolna granica przedziału dominanty. Przedział dominanty to, oczywiście, najliczniej reprezentowany przedział.

Dominantę wyznacza się wtedy, gdy rozkład cech jest *typowy*.

c. kwantyle

Kwantyle to wartości cechy dzielące próbę w określonym stosunku. Na przykład mediana dzieli próbę w stosunku 1:1 – 50% obserwacji posiada wartość cechy nie większą niż mediana i 50% obserwacji posiada wartość nie mniejszą niż ona.

Ogólnie kwantyl rzędu $\frac{b}{v}$ zapisujemy $Q_{\frac{b}{v}}(x)$, gdzie $\frac{b}{v}$ jest sposobem podziału próby przez dany kwantyl. Na przykład $Q_{\frac{2}{4}}(x) \equiv Q_{\frac{1}{2}}(x)$ to właśnie mediana, oznaczana też $M(x)$.

i. szereg rozdzielczy przedziałowy:

Pozycja kwantyla:

$$P_{Q_{\frac{b}{v}}} = \frac{b}{v} \cdot n$$

Wzór na kwantyl:

$$Q_{\frac{b}{v}}(x) = x_{0q} + \left(P_{Q_{\frac{b}{v}}} - n_{q-1}^{sk} \right) \frac{h_q}{n_q}$$

gdzie n_{q-1}^{sk} to liczebność skumulowana przedziału poprzedzającego przedział kwantyla, h_q to rozpiętość przedziału kwantyla, n_q - jego liczebność, zaś x_{0q} - dolna granica przedziału kwantyla.

Bardzo często wyznaczanymi kwantylami są tzw. *kwartyle*: $Q_{1/4}(x)$, $Q_{2/4}(x) = M(x)$ i $Q_{3/4}(x)$. Przykładowo, trzeci kwartył, $Q_{3/4}(x)$, dzieli próbę na 75% obserwacji posiadających wartość cechy nie większą niż trzeci kwartył oraz 25% posiadających wartość nie mniejszą niż on.

2. Miary zróżnicowania (rozproszenia, dyspersji)

a. wariancja:

- i. szereg szczegółowy: $S^2(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2$
- ii. szereg rozdzielczy punktowy: $S^2(x) = \frac{1}{n} \sum_{i=1}^p n_i \cdot (x_i - \bar{X})^2$
- iii. szereg rozdzielczy przedziałowy: $S^2(x) = \frac{1}{n} \sum_{i=1}^k n_i \cdot \left(\overset{o}{x_i} - \bar{X} \right)^2$

b. odchylenie standardowe:

$S(x) = \sqrt{S^2(x)}$ - mówi o ile jednostek wartości cechy przeciętnie różnią się od jej średniej

c. odchylenie przeciętne:

- i. szereg szczegółowy: $d(x) = \frac{\sum_{i=1}^n |x_i - \bar{X}|}{n}$
- ii. szereg rozdzielczy punktowy: $d(x) = \frac{\sum_{i=1}^p n_i \cdot |x_i - \bar{X}|}{n}$
- iii. szereg rozdzielczy przedziałowy: $d(x) = \frac{\sum_{i=1}^k n_i \cdot \left| \overset{o}{x_i} - \bar{X} \right|}{n}$

d. klasyczny współczynnik zmienności:

$$V(x) = \frac{S(x)}{\bar{X}} \quad \text{lub alternatywnie: } V(x) = \frac{d(x)}{\bar{X}}$$

Ponieważ $V(x) \in [0, 1]$, to przyjęło się, że:

$V(x) \leq 0,33$ oznacza słabe zróżnicowanie cechy,

$V(x) \in (0,33; 0,67)$ oznacza umiarkowane zróżnicowanie cechy, zaś

$V(x) \geq 0,67$ oznacza silne zróżnicowanie cechy.

e. rozstęp:

- i. szereg szczegółowy i rozdzielczy punktowy: $R(x) = x_{\max} - x_{\min}$
- ii. szereg rozdzielczy przedziałowy: $R(x) = x_{1k} - x_{01}$, gdzie x_{01} oznacza początek pierwszego przedziału klasowego, zaś x_{1k} - koniec ostatniego.

f. odchylenie ćwiartkowe:

$$Q(x) = \frac{Q_{3/4}(x) - Q_{1/4}(x)}{2}$$

g. pozycyjny współczynnik zmienności:

$$V_p(x) = \frac{Q(x)}{M(x)}_{[0,1]} - \text{interpretacja wyniku jest taka sama, jak w przypadku klasycznego współczynnika zmienności.}$$

3. Miary asymetrii (skośności)

a. absolutna miara asymetrii (trzeci moment centralny):

i. szereg szczegółowy: $\mu_3(x) = \frac{\sum_{i=1}^n (x_i - \bar{X})^3}{n}$

ii. szereg rozdzielczy punktowy: $\mu_3(x) = \frac{\sum_{i=1}^p n_i \cdot (x_i - \bar{X})^3}{n}$

iii. szereg rozdzielczy przedziałowy: $\mu_3(x) = \frac{\sum_{i=1}^k n_i \cdot \left(\overset{o}{x_i} - \bar{X} \right)^3}{n}$

Jeśli:

$\mu_3(x) = 0$, to rozkład cechy jest symetryczny,

$\mu_3(x) > 0$, to rozkład cechy jest asymetryczny prawostronnie (prawoskośny),

$\mu_3(x) < 0$, to rozkład cechy jest asymetryczny lewostronnie (lewoskośny).

b. klasyczny współczynnik asymetrii:

$$A(x) = \frac{\mu_3(x)}{S^3(x)}_{[-2, 2]}$$

Na rodzaj asymetrii wskazuje znak trzeciego momentu centralnego (patrz wyżej). Dodatkowo, przyjmuje się, że:

$|A(x)| \leq 0,7$ oznacza słabą asymetrię,

$|A(x)| \in (0,7; 1,4)$ oznacza asymetrię umiarkowaną, zaś

$|A(x)| \geq 1,4$ oznacza asymetrię silną.

c. współczynnik skośności:

$$A_s(x) = \frac{\bar{X} - D(x)}{S(x)}$$

[-1, 1]

Na rodzaj asymetrii wskazuje znak licznika ($\bar{X} - D(x) > 0$ świadczy o asymetrii prawostronnej – proszę sobie wyobrazić taki przypadek i go przeanalizować). Dodatkowo, przyjmuje się, że:

$|A_s(x)| \leq 0,33$ oznacza słabą asymetrię,

$|A_s(x)| \in (0,33; 0,67)$ oznacza asymetrię umiarkowaną, zaś

$|A_s(x)| \geq 0,67$ oznacza asymetrię silną.

d. pozycyjny współczynnik asymetrii:

$$A_p(x) = \frac{[Q_{3/4}(x) - M(x)] - [M(x) - Q_{1/4}(x)]}{2 \cdot Q(x)}$$

[-1, 1]

Wartość i znak pozycyjnego współczynnika asymetrii interpretujemy tak samo jak wartość współczynnika skośności:

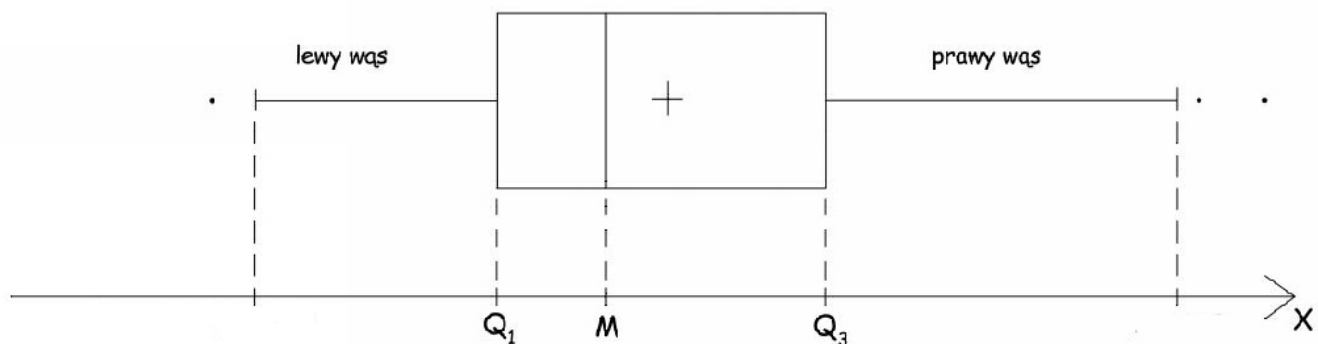
$|A_p(x)| \leq 0,33$ oznacza słabą asymetrię,

$|A_p(x)| \in (0,33; 0,67)$ oznacza asymetrię umiarkowaną, zaś

$|A_p(x)| \geq 0,67$ oznacza asymetrię silną.

Wykres pudełkowy (ramkowy)

Wykres ten służy obrazowaniu rozkładu analizowanej cechy:



Brzośy „pudełka” to odpowiednio pierwszy i trzeci kwartyl. Linia w środku przedstawia medianę, zaś dodatkowy symbol (tutaj „+”) prezentuje wartość średnią. Zróżnicowanie cechy można ocenić po szerokości „pudełka” oraz po długości tzw. wąsów, z których lewy kończy się w punkcie

$$\max\left(x_{\min}, Q_{1/4}(x) - 3 \cdot Q(x)\right), \text{ zaś prawy w punkcie } \min\left(x_{\max}, Q_{3/4}(x) + 3 \cdot Q(x)\right).$$

W niektórych opracowaniach, końcówki wąsów wyznacza się biorąc półtora odchylenia ćwiartkowego, zamiast trzech (jak w powyższym wzorze). Jeszcze inny sposób każe przyjmować koniec lewego wąsa w $x_{\min}$, zaś prawego w $x_{\max}$.

Dodatkowo, z relacji pomiędzy kwartylami można wnioskować o (a)symetrii rozkładu cechy (np. mediana przesunięta w stronę pierwszego kwartyla wskazuje na asymetrię prawostronną). O (a)symetrii

można też wnioskować z długości wąsów (np. odpowiednio dłuższy lewy wąs świadczy o rozciągniętym lewym ramieniu rozkładu cechy).

Wartości cechy „wypadające” poza wąsy, to tzw. *wartości nietypowe* – odbiegające od rozkładu cechy.