Statystyka matematyczna - wykład pierwszy*
Miary parametrów statystycznych – wzory.
kierunek: informatyka i ekonometria I°

dr Jarosław Kotowicz

Spis treści

1 Miary położenia

1.1 Klasyczne miary położenia.

Średnia arytmetyczna

Dla szeregów szczegółowych – średnia arytmetyczna (nieważona, prosta)

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Dla szeregów rozdzielczych punktowych i przedziałowych – średnia arytmetyczna ważona

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{k} x_i n_i \quad \text{(dla szeregu punktowego)}$$

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{k} \dot{x}_i n_i \quad \text{(dla szeregu przedziałowego)},$$

gdzie k jest liczbą klas, a $\dot{x}_i$ środkiem i - tego przedziału klasowego.

Uwaga 1. *1. Jeżeli znamy średnie w każdej z klas tzn. $\overline{x_i}$, gdzie $i \in \overline{1,k}$, to dla szeregu rozdzielczego przedziałowego można liczyć średnią ważoną zgodnie z następującym wzorem*

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{k} \overline{x}_i n_i.$$

2. Od tej chwili nie będziemy rozróżniać szeregów rozdzielczych punktowych i przedziałowych i $\dot{x}_i$ będzie oznaczać dla nich albo środek przedziału, albo wartość punktu.

Średnia harmoniczna

$$\overline{x}_H := \frac{n}{\sum\limits_{i=1}^{n} \frac{1}{x_i}} \quad \text{(dla szeregu szczegółowego)},$$

$$\overline{x}_H := \frac{\sum\limits_{i=1}^{k} n_i}{\sum\limits_{i=1}^{k} \frac{n_i}{\dot{x}_i}} \quad \text{(dla szeregu rozdzielczego)}.$$

Uwaga 2. *Średnia harmoniczną stosujemy wtedy, gdy wartości cechy są podane w przeliczeniu na stałą jednostkę, czyli w postaci wskaźników natężenia, wagi natomiast w jednostkach liczników tych cech.*

Średnia geometryczna

$$\overline{x}_G := \sqrt[n]{\prod_{i=1}^{n} x_i} \quad \text{(dla szeregu szczegółowego)},$$

$$\overline{x}_G := \sqrt[n]{\prod_{i=1}^{k} \dot{x}_i^{n_i}} \quad \text{(dla szeregu rozdzielczego)},$$

gdzie k jest ilością klas, a $n = \sum\limits_{i=1}^{k} n_i$.

Uwaga 3. *Średnia harmoniczną stosujemy przy badaniu średniego tempa zmian zjawiska.*

Uwaga 4. *Należy pamiętać, że dla każdego przypadku powinno się obliczać tylko jedną średnią klasyczną, bowiem tylko jedna jest odpowiednia, a inne tracą sens.*

1.2 Pozycyjne miary położenia

Moda (modalna, dominanta)

Moda jest to wartość cechy statystycznej występującej w rozkładzie empirycznym najczęściej (Mo).

Uwaga 5. *Moda musi być pojedynczą wartością. Jeśli w zbiorze obserwacji nie istnieje pojedyncza wartość cechy statystycznej występująca najczęściej, to moda nie istnieje w tej zbiorowości.*

Dla szeregu rozdzielczego modalną wyznaczamy według wzoru

$$\text{Mo} := x_{om} + \frac{n_m - n_{m-1}}{(n_m - n_{m-1}) + (n_m - n_{m+1})} h_m, \tag{1}$$

gdzie

n_i liczebność przedziału i - tego ($i = m$ zawierającego modalną, $i = m-1$ ($i = m+1$) poprzedzającej przedział (następującym po przedziale) z modalną,

x_{0m} – granica dolna przedziału, w którym znajduje się modalna,

h_m – rozpiętość przedziału, w którym znajduje się modalna.

Mediana

Definicja 1. *Mediana jest to wartość cechy statystycznej, dzieląca zbiór obserwacji na dwie liczebnie równe części (zbiór obserwacji o wartościach mniejszych lub równych oraz zbiór obserwacji o wartościach większych lub równych od wartości mediany).*

$$\text{Me} := \begin{cases} x_{\frac{n+1}{2}} & \text{dla } n \notin \mathbb{N} \setminus 2\mathbb{N} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) & \text{dla } n \in 2\mathbb{N} \end{cases} \quad \text{(dla szeregu szczegółowego)},$$

$$\text{Me} := x_{0\,\text{Me}} + \frac{N_{\text{Me}} - \sum\limits_{i=1}^{m-1} n_i}{n_m} h_m \quad \text{(dla szeregu rozdzielczego)},$$

gdzie
m – numer klasy w której znajduje się mediana,
$x_{0\,\mathrm{Me}}$ – granica dolna przedziału w którym znajduje się mediana,
n_m – liczebność przedziału mediany,
$\sum_{i=1}^{m-1} n_i$ – liczebność skumulowana,
h_m – rozpiętość przedziału mediany,
N_{Me} – pozycja mediany (przyjmuje się, że $N_{\mathrm{Me}} = \frac{n}{2}$).

Kwartyle.

Definicja 2. *Kwartyl i-ty ($i = 1, 2, 3$), jest to wartość cechy statystycznej, dzieląca zbiór obserwacji na dwie części w następujący sposób: w pierwszym zbiorze o liczebności równej co najmniej $\frac{i}{4}$ wszystkich obserwacji znajdują się obserwacje, których wartości nie przekraczają wartości i-tego kwartyla, a w drugim, o liczebności co najmniej $\frac{4-i}{4}$, mamy obserwacje, których wartości są co najmniej równe wartości i-tego kwartyla.*

Uwaga 6. *Kwartyl drugi, to mediana.*

Możliwe są co najmniej dwa sposoby wyznaczania kwartyli dla szeregu szczegółowego

1. Zbiorowość dzielimy na dwie części. W pierwszej są te jednostki których cechy przyjmują wartości mniejsze niż mediana, a w drugiej pozostałe i dla nich wyznaczamy mediany, które będą odpowiednio kwartylem pierwszym Q_1 i trzecim Q_3.

2. Stosujemy następujące wzory

$$Q_1 := \begin{cases} \frac{1}{2}\left(x_{\frac{n}{4}} + x_{\frac{n}{4}+1}\right) & \text{dla } n \in \{m \in \mathbb{N} : 4|m\} \\ x_{\frac{n+1}{4}} & \text{dla } n \in \{m \in \mathbb{N} : 4|m+1\} \\ x_{\frac{n}{4}+\frac{1}{2}} & \text{dla } n \in \{m \in \mathbb{N} : 4|m+2\} \\ \frac{1}{2}\left(x_{\frac{n+1}{4}-\frac{1}{2}} + x_{\frac{n+1}{4}+\frac{1}{2}}\right) & \text{dla } n \in \{m \in \mathbb{N} : 4|m+3\} \end{cases},$$

$$Q_3 := \begin{cases} \frac{1}{2}\left(x_{\frac{3n}{4}} + x_{\frac{3n}{4}+1}\right) & \text{dla } n \in \{m \in \mathbb{N} : 4|m\} \\ x_{\frac{3(n+1)}{4}} & \text{dla } n \in \{m \in \mathbb{N} : 4|m+1\} \\ x_{\frac{3n}{4}+\frac{1}{2}} & \text{dla } n \in \{m \in \mathbb{N} : 4|m+2\} \\ \frac{1}{2}\left(x_{\frac{3(n+1)}{4}-\frac{1}{2}} + x_{\frac{3(n+1)}{4}+\frac{1}{2}}\right) & \text{dla } n \in \{m \in \mathbb{N} : 4|m+3\} \end{cases}.$$

Dla kwartyli pierwszego i trzeciego dla szeregu rozdzielczego stosujemy następujące wzór

$$Q_i := x_{0Q_i} + \frac{N_{Q_i} - \sum_{i=1}^{m-1} n_i}{n_m} h_m,$$

gdzie
m – numer klasy, w której znajduje się kwartyl Q_i,
x_{0Q_i} – granica dolna przedziału, w którym znajduje się kwartyl Q_i,
n_m – liczebność przedziału kwartyla Q_i,
$\sum_{i=1}^{m-1} n_i$ – liczebność skumulowana,
h_m – rozpiętość przedziału kwartyla Q_i,
N_{Q_i} – pozycja kwartyla Q_i (przyjmujemy $N_{Q_1} = \frac{n}{4}$ i $N_{Q_3} = \frac{3n}{4}$).

Kwantyle.

Definicja 3. *Niech $p \in]0, 1[$. Kwantylem rzędu p jednowymiarowej zmiennej losowej X nazywamy liczbę κ_p taką, że*

$$P\left(\{\omega : X(\omega) \leqslant \kappa_p\}\right) \geqslant p \wedge P\left(\{\omega : X(\omega) \geqslant \kappa_p\}\right) \geqslant 1 - p. \tag{2}$$

Będziemy korzystać z następujących wzorów interpolacyjnych dla kwantyli dla szeregu rozdzielczego przedziałowego

$$\kappa_p := x_{0p} + [n(p) - n(x_{0p})]\,\frac{h_p}{n_p}, \tag{3}$$

gdzie
x_{0p} – granica dolna przedziału, w którym znajduje się kwantyl rzędu p,
$n(p)$ – pozycja kwantyla rzędu p (zauważmy, że $n(p) = np$),
$n(x_{0p})$ – liczebność skumulowana dla przedziału poprzedzającego przedział kwantyla rzędu p,
h_p – rozpiętość przedziału, w którym znajduje się kwantyl rzędu p,
n_p – liczebność przedziału, w którym znajduje się kwantyl rzędu p.

2 Miary zmienności

2.1 Klasyczne miary zmienności

Definicja 4. *Wariancją nazywamy liczbę równą*

$$s^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad \textit{(dla szeregu szczegółowego)},$$

$$s^2 := \frac{1}{n}\sum_{i=1}^{k}(\dot{x}_i - \overline{x})^2 n_i \quad \textit{(dla szeregu rozdzielczego)}.$$

Uwaga 7. *Wariancję można też zdefiniować następująco:*

$$\tilde{s}^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad \textit{(dla szeregu szczegółowego)},$$

$$\tilde{s}^2 := \frac{1}{n-1}\sum_{i=1}^{k}(\dot{x}_i - \overline{x})^2 n_i \quad \textit{(dla szeregu rozdzielczego)}.$$

- s^2 nazywana jest wariancją populacyjną, $\tilde{s}^2$ wariancją próbkową[1].

Definicja 5. *Odchylenie standardowe to pierwiastek kwadratowy z wariancji (oznaczamy je s lub $\tilde{s}$ w zależności od tego jak wyznaczamy wariancję).*
Odchylenie przeciętne definiujemy następująco

$$d := \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}| \quad \textit{(dla szeregu szczegółowego)},$$

$$d := \frac{1}{n}\sum_{i=1}^{k}|\dot{x}_i - \overline{x}|n_i \quad \textit{(dla szeregu rozdzielczego)}.$$

Definicja 6. *Klasycznymi współczynnikami zmienności (Pearsona) (dla $\overline{x} \neq 0$) nazywamy liczbę równą odpowiednio*

$$V_s := \frac{s}{\overline{x}}, \quad V_d := \frac{d}{\overline{x}}.$$

Uwaga 8. *Typowy obszar zmienności ma największe zastosowanie w przypadku, gdy dane są wyraźnie skupione wokół wartości średniej.*

Klasyczny typowy obszar zmienności wyznacza się z wykorzystaniem odchylenia standardowego i odchylenia przeciętnego.

Definicja 7. *Typowym obszarem zmienności jest przedziałem określonym warunkiem*

- $T_s :=]\overline{x} - s, \overline{x} + s[$ *(odpowiednio* $]\overline{x} - \tilde{s}, \overline{x} + \tilde{s}[$*)*,
- $T_d :=]\overline{x} - d, \overline{x} + d[$.

[1] Zobacz pomoc do MS Excela.

2.2 Pozycyjne miary zmienności

Definicja 8. *Pozycyjnym rozstępem próby nazywamy liczbę*

$$R := x_{\max} - x_{\min},$$

gdzie $x_{\max} := \max_{i\in\overline{1,n}} x_i$ *oraz* $x_{\min} := \min_{i\in\overline{1,n}} x_i$.

Odchyleniem ćwiartkowym nazywamy liczbę

$$Q := \frac{Q_3 - Q_1}{2}.$$

Pozycyjnym typowym obszarem zmienności nazywamy następujący przedział

$$T_Q :=]\,\mathrm{Me} - Q, \mathrm{Me} + Q[.$$

Definicja 9. *Pozycyjnymi współczynnikami zmienności nazywamy liczby równe odpowiednio*

$$V_Q := \frac{Q}{\mathrm{Me}},\ (\mathrm{Me} \neq 0) \quad oraz \quad V_{Q_1,Q_3} := \frac{Q_3 - Q_1}{Q_3 + Q_1}.$$

Uwaga 9. *Współczynnik zmienności informuje o sile rozproszenia.*

3 Miary asymetrii

3.1 Klasyczne miary asymetrii

Trzeci moment centralny

$$\mu_3 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3 \quad \text{(dla szeregu szczegółowego)},$$

$$\mu_3 := \frac{1}{n}\sum_{i=1}^{k}(\dot{x}_i - \overline{x})^3 n_i \quad \text{(dla szeregu rozdzielczego)}.$$

Definicja 10. *Klasycznym współczynnikiem asymetrii nazywamy liczbę*

$$A := \frac{\mu_3}{s^3}.$$

3.2 Pozycyjne miary asymetrii

Definicja 11. *Pozycyjnym wskaźnikiem asymetrii nazywamy liczbę*

$$W_s^Q := (Q_3 - \mathrm{Me}) - (\mathrm{Me} - Q_1).$$

Pozycyjnym współczynnikiem asymetrii nazywamy liczbę

$$A_Q := \frac{(Q_3 - \mathrm{Me}) - (\mathrm{Me} - Q_1)}{2Q}$$

3.3 Mieszane miary asymetrii

Definicja 12. *Wskaźnikiem asymetrii nazywamy liczbę*

$$W_s := \overline{x} - \mathrm{Mo}\,.$$

Mówimy o asymetrii lewostronne (odpowiednio prawostronnej) wtedy, gdy $W_s < 0$ *(odpowiednio* $W_s > 0$*).*

Definicja 13. *Pierwszym współczynnikiem asymetrii Pearsona nazywamy liczbę*

$$A_s := \frac{\overline{x} - \mathrm{Mo}}{s} \quad oraz \quad A_d := \frac{\overline{x} - \mathrm{Mo}}{d}.$$

Drugi współczynnikiem asymetrii Pearsona nazywamy liczbę

$$W_{s,2} := \frac{\overline{x} - \mathrm{Me}}{s} \quad oraz \quad W_{d,2} := \frac{\overline{x} - \mathrm{Me}}{d}.$$

4 Miary koncentracji

4.1 Krzywa Lorenza

Niech obserwacje z_i, gdzie $i \in \overline{1,n}$ spełniają warunki

$$0 \leqslant z_1 \leqslant z_2 \leqslant \ldots \leqslant z_n,$$
$$\sum_{i=1}^{n} z_i > 0.$$

Definicja 14. *Krzywą Lorenza odpowiadającą obserwacjom z_i nazywamy łamaną łączącą kolejne punkty (x_i, y_i) ($i \in \overline{0,n}$), gdzie*

$$x_0 = y_0 = 0,$$
$$x_k := \frac{k}{n},\ y_k := \frac{\sum_{i=1}^{k} z_i}{\sum_{i=1}^{n} z_i} \quad k \in \overline{1,n}.$$

Uwaga 10. *Krzywa Lorenza jest zawarta w kwadracie jednostkowym, a ponadto łączy dolny lewy wierzchołek kwadratu z górnym prawym.*

4.2 Klasyczne miary koncentracji

Klasyczne miary koncentracji

Czwarty moment centralny

$$\mu_4 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4 \quad \text{(dla szeregu szczegółowego)},$$
$$\mu_4 := \frac{1}{n}\sum_{i=1}^{k}(\dot{x}_i - \overline{x})^4 n_i \quad \text{(dla szeregu rozdzielczego)}.$$

Definicja 15. *Współczynnikiem koncentracji (kurtozą) nazywamy liczbę*

$$\gamma_4 := \frac{\mu_4}{(s^2)^2}.$$

- Przyjmuje się, że zbiorowość statystyczna ma rozkład normalny wtedy, gdy $\gamma_4 = 3$.
- Czwarty moment centralny można też zdefiniować wzorem[2]

$$\mu_4' := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^4 \quad \text{(dla szeregu szczegółowego)},$$
$$\mu_4' := \frac{1}{n-1}\sum_{i=1}^{k}(\dot{x}_i - \overline{x})^4 n_i \quad \text{(dla szeregu rozdzielczego)}.$$

I przyjąć następującą definicję współczynnika koncentracji

$$\gamma_4 := \frac{\mu_4'}{(\tilde{s}^2)^2}.$$

Definicja 16. *Mówimy, że rozkład cechy w populacji jest leptokurtyczny (wysmukły) wtedy, gdy $\gamma_4 > 3$.*
Natomiast, gdy $\gamma_4 < 3$, to o rozkładzie cechy mówimy, że jest platokurtyczny (spłaszczony).

Definicja 17. *Liczbę $K := \gamma_4 - 3$ nazywamy współczynnikiem ekscesu.*

Uwaga 11. *W literaturze czasami współczynnik ekscesu nazywany też jest kurtozą.*

[2] Zobacz [?]

Definicja 18. *Współczynnik Giniego dla obserwacji $x_1, \ldots, x_n$ definiujemy następująco*

$$G := \frac{\sum_{i,j=1}^{n} |x_i - x_j|}{n^2 \overline{x}}. \tag{4}$$

Jeżeli obserwacje x_i, dla $i \in \overline{1,n}$, uporządkowane są rosnąco, to współczynnik Giniego wyraża się wzorem

$$G := \frac{\sum_{i=1}^{n} (2i - n - 1)x_i}{n^2 \overline{x}}.$$

Uwaga 12. *Współczynnik Giniego jest to pole zwarte pomiędzy krzywą Lorenza, a przekątna kwadratu jednostkowego.*

4.3 Pozycyjne miary koncentracji

Definicja 19. *Pozycyjnym współczynnikiem koncentracji nazywamy liczbę*

$$W_s := \frac{D_9 - D_1}{Q_3 - Q_1},$$

gdzie D_i jest i-tym decylem oraz $D_1 = x_{\frac{n}{10}}, D_9 = x_{\frac{9n}{10}}$ dla szeregu szczegółowego i $D_i := x_{0D_i} + \frac{N_{D_i} - \sum_{i=1}^{m-1} n_i}{n_m} h_m$ dla szeregu rozdzielczego, a pozostałe oznaczenia są analogiczne.

Literatura

Literatura

[1] J. Jóźwiak and J. Podgórski. *Statystyka od podstaw.* PWE, Warszawa, piąte, zmienione edition, 2000.