STATYSTYKA OPISOWA

Dr Alina Gleska

Instytut Matematyki WE PP

28 września 2018

MIARY ZMIENNOŚCI (inaczej: rozproszenia, rozrzutu, zróżnicowania, dyspersji) informują o zróżnicowaniu jednostek zbiorowości ze względu na wartość badanej cechy. Podobnie jak miary położenia dzielimy je na

- miary pozycyjne - wyznaczane za pomocą wartości cechy jednostek zajmujących wyróżnioną pozycję,
- miary klasyczne - obliczane na podstawie wszystkich obserwacji, które określają jak bardzo wartości cechy różnią się od średniej arytmetycznej.

Wszystkie miary zmienności, czyli zarówno klasyczne, jak i pozycyjne, można podzielić ze względu na inne kryterium na

- bezwzględne (inaczej: absolutne), które są wielkościami mianowanymi (tzn. mają miana, czyli jednostki miary takie, jak badana cecha),
- względne (inaczej: relatywne), które są wielkościami niemianowanymi lub mogą być wyrażone w procentach.

W związku z powyższym jedynie miary względne nadają się do porównania dyspersji cech o różnych mianach.

ROZSTĘP (inaczej: amplituda wahań, empiryczny zakres zmienności) informuje o zakresie danych. Wzór:

$$R = x_{max} - x_{min}.$$

Miara prosta do obliczenia (zaleta), zależy jednak od dwóch skrajnych pomiarów (wada). Stosowana tylko przy wstępnej analizie danych.

ROZSTĘP MIĘDZYKWARTYLOWY jest to różnica pomiędzy trzecim i pierwszym kwartylem. Wzór:

$$R_0 = Q_3 - Q_1.$$

Rozstęp międzykwartylowy mierzy zakres zmienności 50% środkowych jednostek pozostałych po odrzuceniu 25% najniższych i 25% najwyższych pomiarów.

ODCHYLENIE ĆWIARTKOWE jest to połowa rozstępu międzykwartylowego. Wzór:

$$Q = \frac{R_0}{2} = \frac{Q_3 - Q_1}{2}.$$

Odchylenie ćwiartkowe informuje, jakie jest przeciętne odchylenie 50% środkowych jednostek od mediany i stosuje się je przede wszystkim wtedy, gdy rozkład cechy jest skrajnie asymetryczny.

POZYCYJNY TYPOWY OBSZAR ZMIENNOŚCI CECHY
Za jednostki typowe w danej zbiorowości uznaje się te, których wartości cechy mieszczą się w przedziale $(Me - Q, Me + Q)$, czyli takie, które odchylają się od mediany mniej niż o odchylenie ćwiartkowe.
UWAGA: typowego obszaru zmienności nie należy mylić z dominantą, jak również w przypadku szeregów rozdzielczych przedziałowych z przedziałem zawierającym dominantę. Są to dwa różne pojęcia.

ODCHYLENIE PRZECIĘTNE pozwala określić, jak bardzo poszczególne obserwacje różnią się od średniej arytmetycznej.
Wzór:

$$d = \frac{1}{n} \sum_{i=1}^{n} |x_i - \overline{x}|,$$

gdzie x_i to wartość cechy i-tej jednostki statystycznej (obserwacji), $\overline{x}$ to średnia arytmetyczna, a n to liczebność zbiorowości.
Odchylenie przeciętne interpretuje się jako średni rozrzut pomiarów wokół średniej arytmetycznej.

WŁASNOŚCI:

- odchylenie przeciętne jest zawsze liczbą nieujemną: $d \geq 0$,
- $d = 0$ w przypadku braku zróżnicowania cechy (tj. gdy wszystkie pomiary są takie same),
- im WIĘKSZA jest WARTOŚĆ odchylenia przeciętnego, tym SILNIEJSZE jest ZRÓŻNICOWANIE zbiorowości ze względu na badaną cechę,
- odchylenie przeciętne jest wielkością MIANOWANĄ (wyrażone jest w takich samych jednostkach miary jak badana cecha).

WARIANCJA mierzy średnie odchylenie kwadratowe od średniej arytmetycznej. Wzór:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

gdzie x_i to wartość cechy i-tej jednostki statystycznej (obserwacji), $\overline{x}$ to średnia arytmetyczna, a n to liczebność zbiorowości.

Wzór ten możemy przekształcić do wygodniejszej do stosowania postaci:

$$s^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \overline{x}^2.$$

WŁASNOŚCI:

- wariancja jest zawsze liczbą nieujemną: $s^2 \geq 0$,
- $s^2 = 0$ w przypadku braku zróżnicowania (tj. gdy wszystkie pomiary są takie same),
- im WIĘKSZA jest WARTOŚĆ wariancji, tym SILNIEJSZE jest ZRÓŻNICOWANIE zbiorowości ze względu na badaną cechę,
- wariancja jest wielkością mianowaną, ale wyrażoną w jednostkach miary badanej cechy podniesionych do kwadratu. Nie daje się zatem intuicyjnie zinterpretować.

ODCHYLENIE STANDARDOWE określa, o jaką wartość średnio poszczególne jednostki zbiorowości różnią się ze względu na wartość badanej cechy od średniej arytmetycznej. Wzór:

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

Odchylenie standardowe wyraża się w naturalnych jednostkach miary.

WŁASNOŚCI:

- odchylenie standardowe jest zawsze liczbą nieujemną: $s \geq 0$,
- $s = 0$ w przypadku braku zróżnicowania (tj. gdy wszystkie pomiary są takie same),
- im WIĘKSZA jest WARTOŚĆ odchylenia standardowego, tym SILNIEJSZE jest ZRÓŻNICOWANIE zbiorowości ze względu na badaną cechę,
- odchylenie standardowe jest wielkością mianowaną, wyrażoną w takich samych jednostkach miary jak badana cecha,
- odchylenie standardowe jest zawsze większe od odchylenia przeciętnego $s > d$,
- pomiędzy odchyleniem standardowym, przeciętnym a ćwiartkowym zachodzi związek: $s > d > Q$.

UWAGA: istnieje również inny sposób obliczania wariancji, zgodnie ze wzorem:

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

Oba sposoby obliczania wariancji są prawidłowe, w zależności od tego, co jest celem naszego badania. Procedura obliczeń z użyciem $n-1$ stosowana jest wtedy, gdy dane pochodzą z próby (n<30), a my na podstawie wariancji obliczonej dla tych danych chcemy ocenić wariancję całej populacji. Udowodniono, że wariancja uzyskana w ten nieco sztuczny sposób lepiej przybliża nieznaną wariancję całej populacji niż wariancja, która ma w mianowniku n.

KLASYCZNY TYPOWY OBSZAR ZMIENNOŚCI CECHY
Za jednostki typowe w danej zbiorowości uznaje się te, których wartości cechy mieszczą się w przedziale $(\overline{x} - s, \overline{x} + s)$, czyli takie, które odchylają się od średniej arytmetycznej mniej niż o odchylenie standardowe.
UWAGA: typowego obszaru zmienności nie należy mylić z dominantą, jak również w przypadku szeregów rozdzielczych przedziałowych z przedziałem zawierającym dominantę. Są to dwa różne pojęcia.
Przykład.

Współczynniki zmienności (pozycyjny lub klasyczny) są względnymi miarami zróżnicowania, które są niezależne od jednostek miary. Wykorzystujemy je, gdy chcemy:

- ocenić siłę zróżnicowania analizowanej cechy,
- porównać rozproszenie różnych cech w tej samej zbiorowości,
- porównać rozproszenie tej samej cechy w różnych zbiorowościach.

Współczynniki zmienności najczęściej wyrażane są w procentach, czyli po pomnożeniu przez 100%. Im wartość współczynnika zmienności jest większa, tym bardziej zróżnicowana jest zbiorowość ze względu na badaną cechę. Z kolei wartość tego współczynnika bliska zeru świadczy o bardzo małym zróżnicowaniu zbiorowości, czyli zbiorowość ze względu na badaną cechę jest jednorodna.

POZYCYJNY WSPÓŁCZYNNIK ZMIENNOŚCI określa, jaki procent mediany stanowi odchylenie ćwiartkowe. Wzór:

$$V_Q = \frac{Q}{Me} \cdot 100\%.$$

WŁASNOŚCI:

- $V_Q \geq 0\%$,
- $V_Q = 0\%$, jeżeli zbiorowość nie jest w ogóle zróżnicowana,
- wyższa wartość współczynnika zmienności świadczy o silniejszym zróżnicowaniu cechy, czyli o niejednorodności zbiorowości.

Można przyjąć umowne przedziały współczynnika zmienności V_Q:

- $0\% - 20\%$ – zróżnicowanie cechy słabe,
- $20\% - 40\%$ – zróżnicowanie cechy umiarkowane,
- $40\% - 60\%$ – zróżnicowanie cechy silne,
- powyżej 60% – zróżnicowanie cechy bardzo silne.

KLASYCZNY WSPÓŁCZYNNIK ZMIENNOŚCI określa, jaki procent średniej arytmetycznej stanowi odchylenie standardowe. Wzór:

$$V_s = \frac{s}{\overline{x}} \cdot 100\%.$$

WŁASNOŚCI:

- $V_s \geq 0\%$,
- $V_s = 0\%$, jeżeli zbiorowość nie jest w ogóle zróżnicowana,
- wyższa wartość współczynnika zmienności świadczy o silniejszym zróżnicowaniu cechy, czyli o niejednorodności zbiorowości.

Można przyjąć umowne przedziały współczynnika zmienności V_s:

- $0\% - 20\%$ – zróżnicowanie cechy słabe,
- $20\% - 40\%$ – zróżnicowanie cechy umiarkowane,
- $40\% - 60\%$ – zróżnicowanie cechy silne,
- powyżej 60% – zróżnicowanie cechy bardzo silne.

Wariancja i odchylenie standardowe w szeregu rozdzielczym punktowym:

$$s^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i \quad \text{lub} \quad s^2 = \left(\frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i\right) - \overline{x}^2,$$

lub ze wskaźnikami struktury:

$$s^2 = \sum_{i=1}^{k}(x_i - \overline{x})^2 \omega_i \quad \text{lub} \quad s^2 = \left(\sum_{i=1}^{k} x_i^2 \omega_i\right) - \overline{x}^2,$$

bądź z udziałami procentowymi:

$$s^2 = \frac{1}{100}\sum_{i=1}^{k}(x_i - \overline{x})^2 p_i \quad \text{lub} \quad s^2 = \left(\frac{1}{100}\sum_{i=1}^{k} x_i^2 p_i\right) - \overline{x}^2.$$

Wariancja i odchylenie standardowe w szeregu rozdzielczym przedziałowym:

$$s^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i^0 - \overline{x})^2 n_i \quad \text{lub} \quad s^2 = \left(\frac{1}{n}\sum_{i=1}^{k}(x_i^0)^2 n_i\right) - \overline{x}^2,$$

lub ze wskaźnikami struktury:

$$s^2 = \sum_{i=1}^{k}(x_i^0 - \overline{x})^2 \omega_i \quad \text{lub} \quad s^2 = \left(\sum_{i=1}^{k}(x_i^0)^2 \omega_i\right) - \overline{x}^2,$$

bądź z udziałami procentowymi:

$$s^2 = \frac{1}{100}\sum_{i=1}^{k}(x_i^0 - \overline{x})^2 p_i \quad \text{lub} \quad s^2 = \left(\frac{1}{100}\sum_{i=1}^{k}(x_i^0)^2 p_i\right) - \overline{x}^2.$$

RÓWNOŚĆ WARIANCYJNA

Zdarza się, że zbiorowość jest podzielona na pewne grupy, przy czym znamy wariancje poszczególnych grup. Chcielibyśmy na podstawie tych cząstkowych wariancji wyznaczyć wariancję całej zbiorowości, czyli tzw. wariancję ogólną.

Wariancję ogólną wyznacza się na podstawie równości wariancyjnej. Można udowodnić, że wariancja ogólna jest sumą dwóch składników:

- wariancji wewnątrzgrupowej $\overline{s_i^2}$ (jest to średnia ważona wariancji poszczególnych grup),
- wariancji międzygrupowej $s^2(\overline{x_i})$ (jest to wariancja średnich arytmetycznych poszczególnych grup).

Wariancję wewnątrzgrupową obliczamy zgodnie ze wzorem:

$$\overline{s_i^2} = \frac{1}{n}\sum_{i=1}^{k} s_i^2 \cdot n_i,$$

gdzie s_i^2 - wariancja i-tej grupy, n_i - liczebność i-tej grupy, k - liczba grup, n - liczebność zbiorowości.
Wariancję międzygrupową obliczamy zgodnie ze wzorem:

$$s^2(\overline{x_i}) = \frac{1}{n}\sum_{i=1}^{k} (\overline{x_i} - \overline{\overline{x}})^2 \cdot n_i,$$

gdzie $\overline{x_i}$ - średnia arytmetyczna i-tej grupy, $\overline{\overline{x}}$ - średnia arytmetyczna całej zbiorowości, obliczana jako średnia ważona średnich dla grup.

Zatem całościowy wzór na obliczenie wariancji ogólnej:

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} s_i^2 \cdot n_i + \frac{1}{n}\sum_{i=1}^{k} (\overline{x_i} - \overline{\overline{x}})^2 \cdot n_i.$$

Przykład.
Rozkład normalny.

REGUŁA CZEBYSZEWA
W dowolnym rozkładzie co najmniej $1 - \frac{1}{k^2}$ jednostek zbiorowości znajduje się w przedziale $(\overline{x} - ks, \overline{x} + ks)$, przy czym twierdzenie to jest prawdziwe dla $k > 1$. Zatem w przedziale:

- $(\overline{x} - 2s, \overline{x} + 2s)$ leży co najmniej $\frac{3}{4}$, czyli co najmniej 75% jednostek zbiorowości,
- $(\overline{x} - 3s, \overline{x} + 3s)$ leży co najmniej $\frac{8}{9}$, czyli co najmniej 89% jednostek zbiorowości,
- $(\overline{x} - 4s, \overline{x} + 4s)$ leży co najmniej $\frac{15}{16}$, czyli co najmniej 93,75% jednostek zbiorowości,
- $(\overline{x} - 5s, \overline{x} + 5s)$ leży co najmniej $\frac{24}{25}$, czyli co najmniej 96% jednostek zbiorowości.

Przykład zastosowania.