

Wskaźniki położenia i rozproszenia

dr Mariusz Grządziel

Katedra Matematyki, Uniwersytet Przyrodniczy we Wrocławiu

r. akad. 2022/2023

Populacja i próba

Populacja- zbiorowość skończona lub nieskończona, w stosunku do której mają być formułowane wnioski.

Próba- skończony podzbiór populacji podlegający szczegółowemu badaniu.

Rozważane zbiory danych— można je interpretować jako próby z pewnych populacji; w dalszym ciągu-terminy *próba* i *zbiór danych* będą używane zamiennie.

Wskaźniki sumaryczne

Histogram— sugestywny środek syntezy informacji zawartej w zbiorze danych;

wskaźniki sumaryczne — miary liczbowe pozwalające na zwięzły opis zbioru danych — lub populacji (zbiorowości), z której ten zbiór danych został wybrany.

Wskaźniki:

- położenia — określają centrum zbioru danych;
- rozproszenia — określają rozproszenie cechy wokół wskaźnika położenia.

Przykład

Ceny mieszkań używanych we Wrocławiu (ceny proponowane przez sprzedających):

Ceny mieszkań w dzielnicy A (miasta Wrocławia):

65,80,139,180,355,158,240,205,265,305,200,155,209,
310,149,254,188,265,275,200,184,130,260,250,195

Ceny mieszkań w dzielnicy B (miasta Wrocławia):

420,350,275,277,327,223,198,275,350,327,260,306,
270,295,270,220,299

Średnia cena w dzielnicy A: 208,8; średnia cena w dzielnicy B:
290,71.

Czy ceny mieszkań w dzielnicy B są wyższe niż w dzielnicy A?

Wskaźniki położenia

Niech x_1, x_2, \dots, x_n oznacza próbę o licznosci n .

Definicja 1

Wartością średnią w próbie, oznaczaną \bar{x} , nazywamy średnią arytmetyczną

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

Dla danych dotyczących cen mieszkań w dzielnicy B wartość średnia wynosi:

$$\bar{x} = \frac{1}{17}(420 + 350 + \dots + 299) \approx 290,71$$

Mediana

Średnia w próbie — sensowna, gdy histogram jest w przybliżeniu symetryczny (tak jak w przypadku histogramu dla danych dotyczących cen mieszkań w dzielnicy B).

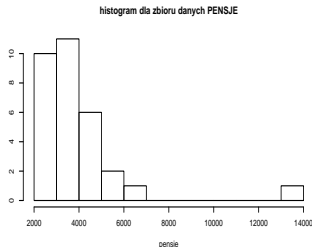
Przykład. Wynagrodzenie pracowników w pewnej firmie:

2400 zł (1 osoba), 2900 zł (9 osób), 3100 zł (6 osób), 3400 zł (5 osób), 4100 zł (4 osoby), 4800 zł (2 osoby), 6000 zł (2 osoby), 6500 zł (1 osoba), 14000 zł (1 osoba). Średnia wynagrodzenie wynosi:

$$\frac{2400+9\times 2900+6\times 3100+5\times 3400+4\times 4100+2\times 4800+2\times 6000+14000}{31} \approx 3954,84$$

Można oczekiwać, że histogram dla tych danych nie będzie symetryczny.

Mediana— c.d.



Rysunek: Histogram dla danych PENSJE

Histogram dla danych PENSJE nie jest symetryczny— ma „prawy ogon” dłuższy niż „lewy ogon”. Nawiązując do terminologii z książki J. Koronackiego i J. Mielniczuka (Rozdz. 1.2) jest on prawostronnie skośny.

Analogicznie określamy lewostronna skośność histogramu.

W tym przypadku bardziej sensownym wskaźnikiem położenia będzie tzw. mediana.

Mediana — definicja

Uporządkowane niemalejąco elementy próby x_1, x_2, \dots, x_n oznaczmy przez

$$x_{(1)}, x_{(2)}, \dots, x_{(n-1)}, x_{(n)},$$

gdzie $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n-1)} \leq x_{(n)}$. Dla danych PENSJE:

$$x_{(1)} = 2400;$$

$$x_{(2)} = x_{(3)} = \dots = x_{(10)} = 2900;$$

$$x_{(11)} = x_{(12)} = \dots = x_{(16)} = 3100;$$

$$x_{(17)} = x_{(18)} = \dots = x_{(21)} = 3400$$

...

$$x_{(30)} = 6500$$

$$x_{(31)} = 14000$$

Definicja 2

Medianą w próbie (lub medianą próby), oznaczaną x_{med} nazywamy następującą wielkość

$$x_{med} = \begin{cases} x_{((n+1)/2)}, & \text{jeśli } n \text{ jest nieparzyste,} \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}), & \text{jeśli } n \text{ jest parzyste.} \end{cases}$$

Cechą mediany jest odporność na **obserwacje odstające**, czyli wartości bardzo wyraźnie odstające od innych obserwacji w próbie.

Jeśli w danych PENSJE 14000 zamienić na 140000, wartość mediany się nie zmieni!

Inne odporne wskaźniki położenia

Średnia ucinana (z parametrem k). Otrzymujemy ją odrzucając k najmniejszych i k największych obserwacji w próbie, a następnie obliczając średnią dla pozostałych elementów próby.

Średnia winsorowska z parametrem k . Otrzymujemy ją zastępując k najmniejszych elementów próby elementem $x_{(k+1)}$ a k największych elementem $x_{(n-k)}$ i obliczając dla tak zmodyfikowanej próby średnią.

Wskaźniki rozproszenia

Definicja 3

Rozstępem próby o licznosci n , oznaczanym przez R , nazywamy różnicę

$$R = x_{(n)} - x_{(1)}$$

gdzie $x_{(1)}$ i $x_{(n)}$ są, odpowiednio, najmniejszym i największym elementem w próbie.

Definicja 4

Wariancję w próbie, oznaczaną przez s^2 , określamy wzorem

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

gdzie \bar{x} oznacza średnią w próbie. Pierwiastek z wariancji nazywamy odchyleniem standardowym w próbie; oznaczamy go przez s .

Kwartyle i rozstęp międzykwartyłowy

Uporządkowane niemalejąco elementy próby x_1, x_2, \dots, x_n oznaczamy przez

$$x_{(1)}, x_{(2)}, \dots, x_{(n-1)}, x_{(n)}.$$

Niech $m = 1 + 0,25 \cdot (n - 1)$.

Definicja 5

Kwartył dolny Q_1 określamy wzorem

$$Q_1 = \begin{cases} x_{(m)}, & \text{jeśli } m \text{ jest całkowite,} \\ (1 - w)x_{(d)} + w \cdot x_{(d+1)}, & \text{jeśli } m \text{ nie jest całkowite,} \end{cases}$$

gdzie $d = E(m)$ (d jest równa części całkowitej z m) a $w = m - d$.

Kwartył górny określamy analogicznie, przyjmując $m = 1 + 0,75 \cdot (n - 1)$.

Definicja 6

Rozstępem międzykwartyłowym IQR nazywamy różnicę

$$IQR = Q_3 - Q_1.$$

Uwagi o innych definicjach kwartyli

W środowisku R kwartyle są obliczane wg. wyżej podanej definicji (jeżeli użyta jest opcja domyślna).

W książce Koronackiego i Mielniczuka (2001) kwartył dolny definiowany jest jako mediana dla podpróby składającej się z obserwacji mniejszych niż mediana; analogicznie jest tam definiowany kwartył górny.

Niektórzy autorzy zamiast kwartyli obliczają tzw. *zawiasy* (ang. *hinges*). Zawiasy dolny i górny będziemy oznaczać przez, odpowiednio, \tilde{Q}_1 i \tilde{Q}_3 .

\tilde{Q}_1 jest „w przybliżeniu” równe Q_1 a \tilde{Q}_3 jest „w przybliżeniu” równe Q_3 .

Zawiasy Tukeya

Uporządkowane niemalejąco elementy próby x_1, x_2, \dots, x_n oznaczmy przez

$$x_{(1)}, x_{(2)}, \dots, x_{(n-1)}, x_{(n)},$$

gdzie $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n-1)} \leq x_{(n)}$. Niech r będzie resztą z dzielenia n przez 4, tj.

$$n = 4m + r \quad m \text{ jest liczbą całkowitą, } r \in \{0, 1, 2, 3\}$$

„Zawiasy” Tukeya \tilde{Q}_1 i \tilde{Q}_3 dla tej próby są zdefiniowane następującymi wzorami:

$$\tilde{Q}_1 = \begin{cases} \frac{1}{2}(x_{(m)} + x_{(m+1)}), & \text{jeśli } r = 0, \\ x_{(m+1)}, & \text{jeśli } r = 1, \\ x_{(m+1)}, & \text{jeśli } r = 2, \\ \frac{1}{2}(x_{(m+1)} + x_{(m+2)}), & \text{jeśli } r = 3, \end{cases}$$
$$\tilde{Q}_3 = \begin{cases} \frac{1}{2}(x_{(3m)} + x_{(3m+1)}), & \text{jeśli } r = 0, \\ x_{(3m+1)}, & \text{jeśli } r = 1, \\ x_{(3m+2)}, & \text{jeśli } r = 2, \\ \frac{1}{2}(x_{(3m+2)} + x_{(3m+3)}), & \text{jeśli } r = 3. \end{cases}$$

Wykres ramkowy

Dla danych dotyczących cen mieszkań w osiedlu B mamy:

$$\begin{aligned} MIN = 198; \quad \tilde{Q}_1 = Q_1 = 270; \quad x_{med} = 277; \\ \tilde{Q}_3 = Q_3 = 327; \quad MAX = 420; \end{aligned}$$

gdzie MIN i MAX oznaczają, odpowiednio, minimalną i maksymalną obserwację w próbie (zbiorze danych).

Wykres ramkowy składa się z „ramki”-dolna podstawa ramki ma współrzędną y -wą równą \tilde{Q}_1 , współrzędna górnej podstawy ramki jest równa \tilde{Q}_3 ; poziomy odcinek wewnątrz ramki odpowiada medianie w próbie.

Wykres ramkowy— c.d.

Określamy $\widetilde{IQR} := \widetilde{Q}_3 - \widetilde{Q}_1$. Odcinek wychodzący z górnej podstawy ramki kończy się linią poziomą o współrzędnej pionowej równej:

$$m_1 = \begin{cases} MAX, & \text{jeśli } MAX \leq \widetilde{Q}_3 + 1,5 \times \widetilde{IQR}, \\ MAX0, & \text{jeśli } MAX > \widetilde{Q}_3 + 1,5 \times \widetilde{IQR}, \end{cases}$$

gdzie $MAX0$ oznacza największą obserwację ze zbioru obserwacji mniejszych lub równych $\widetilde{Q}_3 + 1,5 \times \widetilde{IQR}$.

Jeśli $MAX > \widetilde{Q}_3 + 1,5 \times \widetilde{IQR}$, to zaznaczamy na wykresie wszystkie obserwacje większe niż $\widetilde{Q}_3 + 1,5 \times \widetilde{IQR}$.

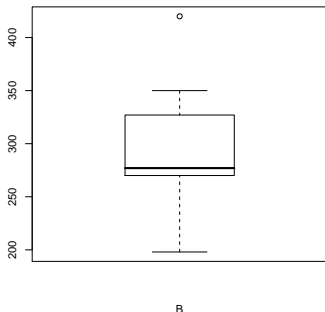
Analogicznie, odcinek wychodzący z dolnej podstawy ramki kończy się linią poziomą o współrzędnej y-wej równej:

$$m_2 = \begin{cases} MIN, & \text{jeśli } MIN \geq \widetilde{Q}_1 - 1,5 \times \widetilde{IQR}, \\ MIN0, & \text{jeśli } MIN < \widetilde{Q}_1 - 1,5 \times \widetilde{IQR}, \end{cases}$$

gdzie $MIN0$ oznacza najmniejszą obserwację ze zbioru obserwacji większych lub równych $\widetilde{Q}_1 - 1,5 \times \widetilde{IQR}$.

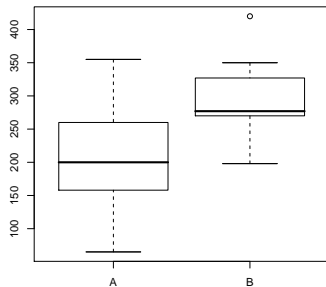
Wykres ramkowy— c.d.

Jeśli $MIN < \tilde{Q}_1 - 1,5 \times \widetilde{IQR}$, to zaznaczamy na wykresie wszystkie obserwacje mniejsze niż $\tilde{Q}_1 - 1,5 \times \widetilde{IQR}$.



Rysunek: Wykres ramkowy dla danych dotyczących cen mieszkań w B

Wykres ramkowe — dla danych z dwóch prób



Rysunek: Wykresy ramkowe dla danych dotyczących cen mieszkań w A i B

Zbiór danych *cabbages*

Zbiór danych *cabbages* (kapusty).

Zawiera pomiary masy i zawartości kwasu askorbinowego dla 60 główek kapusty. Masa jest mierzona w kilogramach; jednostka pomiaru kwasu askorbinowego nie jest wyspecyfikowana.

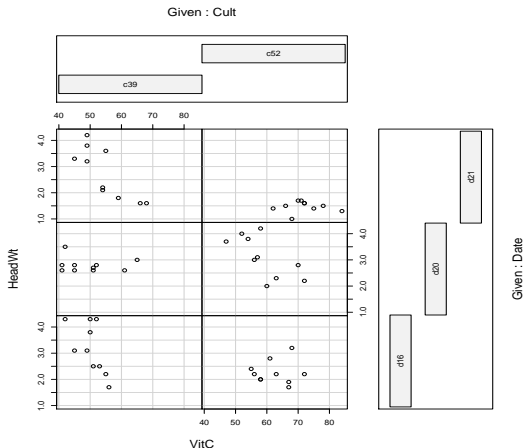
Dwie odmiany kapusty: *c39* i *c52* były zbierane w trzech terminach: *d16*, *d20* i *d21*.

Zbiór danych *cabbages*

```
> library(MASS)
> head(cabbages)
  Cult Date HeadWt VitC
1  c39 d16    2.5   51
2  c39 d16    2.2   55
3  c39 d16    3.1   45
4  c39 d16    4.3   42
5  c39 d16    2.5   53
6  c39 d16    4.3   50
```

Nazwa odmiany — zapisana w kolumnie (zmiennej) *Cult*; *d16*, *d20* i *d21* — kody oznaczające termin zbioru — zapisane są w kolumnie (zmiennej) *Date*.

Zbiór danych *cabbages* przedstawiony graficznie przy użyciu polecenia *cplot*



Wykres został wygenerowany przy użyciu polecenia *cplot(cabbages)*

Zbiór danych *cabbages* — zawartość kwasu askorbinowego

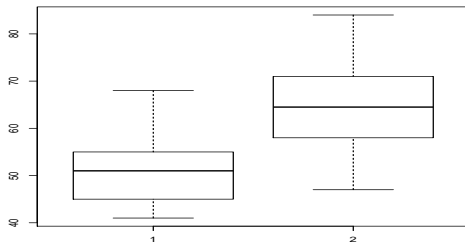
Dane dotyczące zawartości kwasu askorbinowego w główkach kapusty odmiany c39 (uporządkowane niemalejąco):

51 55 45 42 53 50 50 52 56 49 65 52 41 51 41 45 51 45 61 42
54 59 66 54 45 49 49 55 49 68

Dane dotyczące zawartości kwasu askorbinowego w główkach kapusty odmiany c52 (uporządkowane niemalejąco):

58 55 67 61 67 68 58 63 56 72 52 70 57 58 47 56 72 63 54 60
78 75 70 84 71 72 62 68 66 72



Zbiór danych *cabbages* — porównanie zawartości kwasu askorbinowego w odmianach kapusty c39 i c52



Wykres został wygenerowany przy użyciu poleceń:

```
attach(cabbages)
z1=VitC[Cult=="c39"]
z2=VitC[Cult=="c52"]
boxplot(z1, z2)
```

Bibliografia

-  Koronacki, J., Mielniczuk, J. Statystyka dla studentów kierunków technicznych i przyrodniczych. WNT. Warszawa 2001.
-  Łomnicki, A., Wprowadzenie do statystyki dla przyrodników. PWN. Warszawa 2003.