2.Wstępna analiza danych— c.d.- wykład z 3.03.2008

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 popualcji; 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.

Wskaźniki położenia

Niech $x_1, x_2, \ldots, x_n$ oznacza próbę o liczności 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:

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

Mediana

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

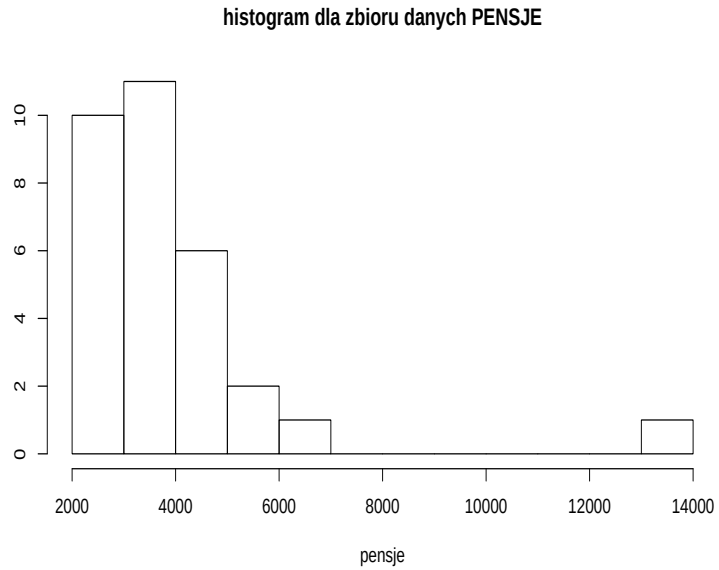
Przykład. Wynagrodznie pracowników w pewnej firmie:

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:

$$\frac{2400+9\times 2900+6\times 3100+5\times 3400+4\times 41002\times 4800+2\times 6000+14000}{31} \approx 3954{,}84$$

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

Mediana— c.d.



Rysunek 1: 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

Uprządkowane niemalejąco elementy próby $x_1, x_2, \ldots, x_n$ oznaczmy przez

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

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

$x_{(1)} = 2400$; $x_{(2)} = x_{(3)} = \ldots = x_{(10)} = 2900$;

$x_{(11)} = x_{(12)} = \ldots = x_{(16)} = 3100$; $x_{(17)} = x_{(18)} = \ldots = x_{(21)} = 3400$

$\ldots$

$x_{(30)} = 6500$ $x_{(31)} = 14000$

Definicja 2 *Medianą w próbie (lub medianą próby), oznaczną* x_{med} *nazywamy następującą wielkość*

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

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 najmiejszych 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 liczności 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 międzykwartylowy

Uprządkowane niemalejąco elementy próby $x_1, x_2, \ldots, x_n$ oznaczamy przez

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

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

Definicja 5 *Kwartyl dolny Q_1 określamy wzorem*

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

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

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

Definicja 6 *Rozstępem międzykwartylowym 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 [1] kwartyl dolny definowany jest jako mediana dla podpróby składającej się z obserwacji mniejszych niż mediana; analogicznie jest tam definiowany kwartyl 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 — dokładne definicje znajdują się na liście zadań nr 2!

Wykres ramkowy

Dla danych dotyczących cen mieszkań w osiedlu B mamy:

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

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

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

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

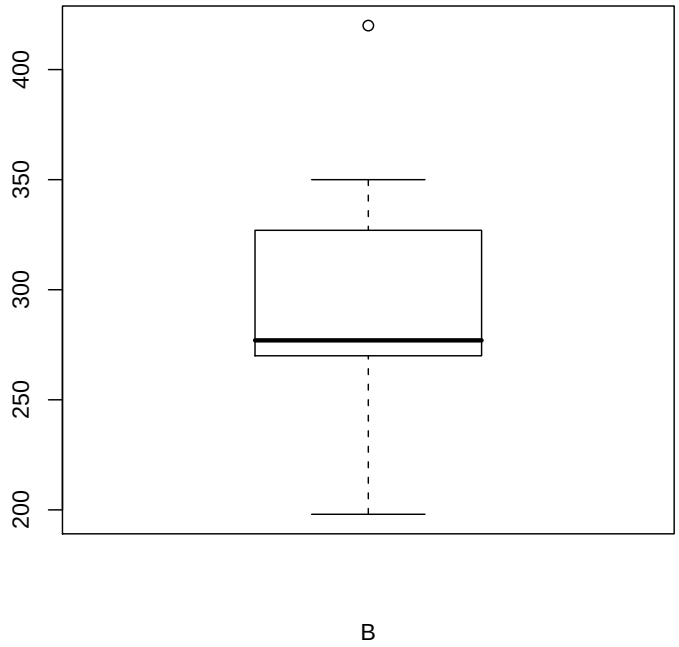
Jeśli $MAX > \tilde{Q}_3 + 1{,}5 \times \widetilde{IQR}$, to zaznaczamy na wykresie wszystkie obserwacje większe niż $\tilde{Q}_3 + 1{,}5 \times \widetilde{IQR}$.

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

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

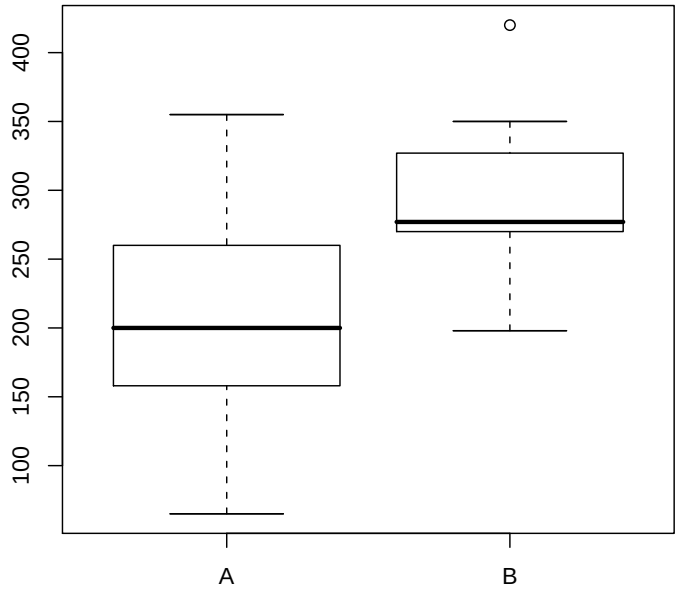
gdzie $MIN0$ oznacza najmniejszą obserwację ze zbioru obserwacji większych lub równych $\tilde{Q}_1 - 1{,}5 \times \widetilde{IQR}$.

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 2: Wykres ramkowy dla danych dotyczących cen mieszkań w B

Wykres ramkowe — dla danych z dwóch prób



Rysunek 3: Wykresy ramkowy dla danych dotyczących cen mieszkań w A i B

Literatura

[1] Koronacki, J., Mielniczuk, J. Statystyka dla studentów kierunków technicznych i przyrodniczych. WNT. Warszawa 2001.