

Wykład 11: Elementy statystyki— c.d.

dr Mariusz Grządziel

20 grudnia 2010

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 $x_1, x_2, \dots, x_n$ 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 NT (dotyczących temperatur zdrowych ludzi):

$$\bar{x} = \frac{1}{130} (96,3 + 96,4 + \dots + 100,8) \approx 98,25$$

Mediana

Średnia w próbie — sensowna, gdy histogram jest w przybliżeniu symetryczny.

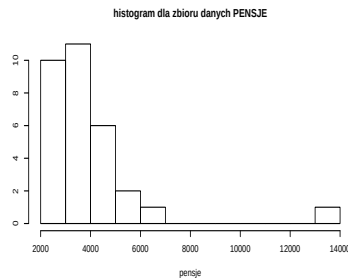
Przykład. Wynagrodzenie pracowników w pewnej firmie:

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.

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 lewostronną skośność histogramu. Dla danych PENSJE bardziej sensownym wskaźnikiem położenia niż średnia jest tzw. mediana.



Rysunek 1: Histogram dla danych PENSJE

Mediana— definicja

Uprządkowane 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:

$$\begin{aligned} x_{(1)} &= 2400; & x_{(2)} &= x_{(3)} = \dots = x_{(10)} = 2900; \\ x_{(11)} &= x_{(12)} = \dots = x_{(16)} = 3100; & x_{(17)} &= x_{(18)} = \dots = x_{(21)} = 2900 \\ & & \dots & \\ x_{(30)} &= 6500 & x_{(31)} &= 14000 \end{aligned}$$

Definicja 2. Medianą w próbie (lub medianą próby), oznaczoną 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!

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 .

Wskaźniki rozproszenia— przykład

Dla danych dotyczących temperatury zdrowych ludzi (zbiór danych NT) rozstęp jest równy:

$$x_{(130)} - x_{(1)} = 100,8 - 96,3 = 4,5,$$

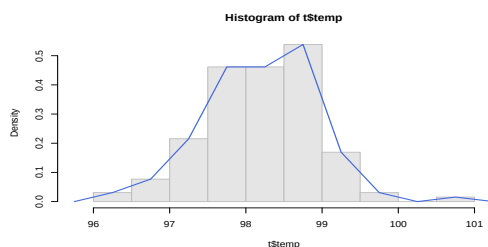
wariancja jest równa:

$$s^2 \approx \frac{1}{129} ((96,2 - 98,25)^2 + (96,3 - 98,25)^2 + \dots + (100,8 - 98,25)^2) \approx 0,538$$

a odchylenie standardowe:

$$s = \sqrt{s^2} = \sqrt{0,538} \approx 0,733.$$

Histogram+wielobok częstotliwości



Rysunek 2: Histogram+wielobok częstotliwości

Problem: znaleźć krzywą gładką przybliżającą histogram— taką, że pole pod nią jest równe 1; krzywa taka odpowiadałaby pewnej funkcji różniczkowalnej i nieujemnej.

Gęstość normalna

Gęstość normalna $\phi_{\mu,\sigma}$ (dla paramerów $\mu \in \mathbb{R}, \sigma > 0$) funkcja określona wzorem:

$$\phi_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Zauważmy, że

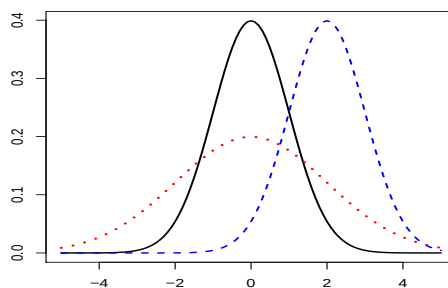
$$\phi_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right),$$

a więc funkcja $\phi_{\mu,\sigma}(x)$ może być przedstawiona jako

$$\phi_{\mu,\sigma}(x) = \frac{1}{\sigma} \phi_{0,1}(g(x)),$$

gdzie $g(x) = \frac{x-\mu}{\sigma}$.

Uwaga. Będziemy w dalszym ciągu używać oznaczenia $\phi(x) = \phi_{0,1}(x)$. Funkcja $\phi(x)$ odpowiada tzw. standardowemu rozkładowi normalnemu.



Rysunek 3: Wykresy gęstości rozkładów normalnych z parametrami $\mu = 0$ i $\sigma = 1$ (linia ciągła), $\mu = 0$ i $\sigma = 2$ (linia „kropkowana”), $\mu = 2$ i $\sigma = 1$ (linia „kreskowana”).

Problem doboru parametrów μ i σ dla konkretnych danych

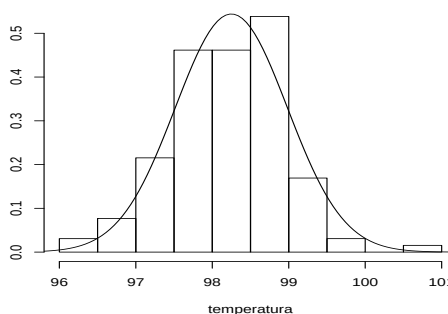
Problem: chcemy wybrać parametry μ i σ tak, aby wykres funkcji $\phi_{\mu,\sigma}(x)$ był możliwie najlepiej „dopasowany” do histogramu probabilistycznego sporządzonego dla danych NT.

W praktyce: odpowiedni histogram probabilistyczny „aproxymujemy” gęstością normalną z parametrami

- μ równemu $\bar{x}$ (średniej z próby);
- σ równemu s (odchyleniu standardowemu z próby).

Dla danych NT: $\bar{x} = 98,25$, $s = 0,733$.

Histogram+ krzywa normalna



Rysunek 4: Histogram probabilistyczny dla danych NT z dołączonym wykresem gęstości normalnej z parametrami $\mu = 98,25$ i $\sigma = 0,733$.