Statystyki Opisowe

Statystyka i analiza danych 2019

Jurek Błaszczyński

10 marca 2019

Miary położenia (tendencji centralnej)

- Dominanta (moda)
- Średnia arytmetyczna (μ, $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$)
- Mediana ($x_{med} = x_{(n+1)/2}$; gdy rozmiar próby n jest nieparzysty)
- Percentyle (k-tym percentylem nazywamy wartość należącą do próby, poniżej której znajduje się k% obserwacji z tej próby)
- Kwartyle (pierwszy kwartyl Q_1 to 25 percentyl, drugi kwartyl Q_2 (mediana) to 50 percentyl, trzeci kwartyl Q_3 to 75 percentyl)
- Średnia geometryczna ($\overline{x}_G = \sqrt[n]{\prod_{i=1}^{n} x_i}$)
- Średnia harmoniczna ($\overline{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$)

Średnia a mediana

- Czy średnia i/lub mediana należą do zbioru wartości przyjmowanych przez dane?
- Która z miar jest bardziej odporna na wartości odstające?
- Według Facebooka (2011), średnia liczba znajomych użytkowników portalu wynosi 190, a połowa użytkowników ma więcej niż 100 znajomych. Dlaczego?
- Domy w pewnej okolicy mają średnią wartość 1mln PLN, ale mediana wynosi tylko 600 tys. PLN. Skąd ta różnica?
- Ile wynosi suma odchyleń od średniej $\sum_{i=1}^{n}(x_i - \overline{x})$?

Miary rozproszenia

- Rozstęp ($R = x_{\mathrm{max}} - x_{\mathrm{min}}$)
- Wariancja (σ^2, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$)
- Odchylenie standardowe (σ, $S = \sqrt{S^2}$)
- Rozstęp międzykwartylowy ($IQR = Q_3 - Q_1$)

Wariancja i odchylenie standardowe

- Jak zmienia się wariancja gdy do wszystkich obserwacji w próbie dodamy stałą?
- Co się stanie z wariancją gdy wszystkie obserwacje w próbie przemnożymy przez stałą?
- Po co nam odchylenie standardowe?

Skośność i kurtoza

- Moment centralny (rzędu k: $M_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k$)
- Współczynnik asymetrii lub skośności ($\frac{M_3}{S^3}$)
- Współczynnik koncentracji lub kurtoza ($\frac{M_4}{S^4}$)

Nierówność Markowa

Załóżmy, że średnia pewnych nieujemnych wartości (liczb) x_i wynosi $\overline{x}$.

Wtedy stosunek wartości większych lub równych od wybranego x' jest mniejszy lub równy $\frac{\overline{x}}{x'}$.

Formalnie
Niech X będzie nieujemną zmienną losową i załóżmy, że istnieje $\mathbb{E}(X)$. Wtedy, dla każdego $x' > 0$:

$$P(X \geq x') \leq \frac{\mathbb{E}(X)}{x'}.$$

Nierówność Czebyszewa

Załóżmy, że średnia pewnych wartości (liczb) x_i wynosi $\overline{x}$ a odchylenie standardowe wynosi S.

Wtedy stosunek wartości oddalonych od średniej $\overline{x}$ o wybrane k lub więcej odchyleń standardowych jest mniejszy lub równy $\frac{1}{k^2}$.

Formalnie

Niech $\mu = \mathbb{E}(X)$, $\sigma^2 = \mathbb{D}^2(X)$ i $Z(X) = \frac{X-\mathbb{E}(X)}{\mathbb{D}(X)} = \frac{X-\mu}{\sigma}$.

Wtedy:

$$P(|X-\mu| \geq x') \leq \frac{\sigma^2}{x'^2},$$

$$P(|Z| \geq k) \leq \frac{1}{k^2}.$$