Statystyki opisowe

SW

Statystyki opisowe

- Miary tendencji centralnej
  - średnia, dominanta
  - percentyle, kwartyle, mediana
- Miary zmienności / rozproszenia
  - rozstęp
  - odstęp międzykwartylowy
  - wariancja i odchylenie standardowe
  - współczynnik zmienności
- Miary asymetrii i spłaszczenia
  - skośność
  - kurtoza

Średnia arytmetyczna

- Średnia w próbie

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

- Średnia w populacji

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Średnia geometryczna

Stosowana w przypadku jednostek multiplikatywnych (np. stopa oprocentowania)

$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Średnia harmoniczna

Stosowana w przypadku jednostek względnych (np. prędkość, gęstość zaludnienia, ...)

$$\overline{x}_H = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{x_i}}$$

Uwaga

Wiele narzędzi stosuje uproszczony wzór na średnią harmoniczną przyjmując jednostkowe wagi w_i zwiazane z poszczególnymi pomiarami x_i.

Percentyle, kwartyle, mediana

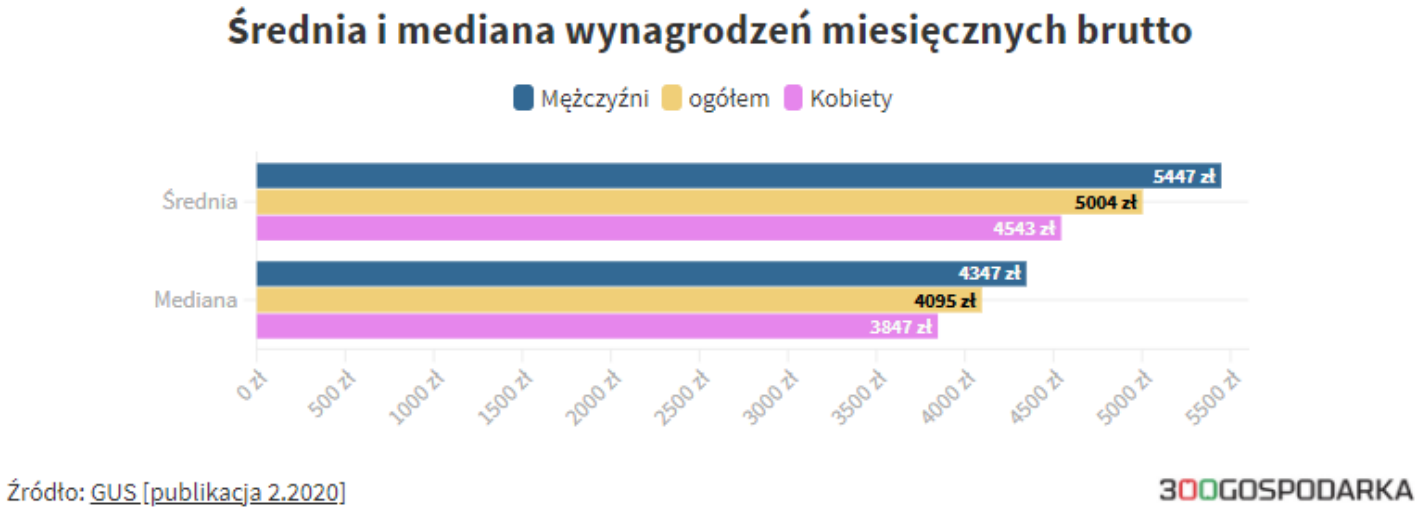
- *P*-ty percentyl w (uporządkowanym rosnąco) zbiorze to taka wartość, poniżej której znajduje się $P\%$ liczb z tego zbioru

$$i_P = (n+1) \times P/100$$

- Często rozważane percentyle
  - $Q1$ = I (*dolny*) kwartyl = 25-ty percentyl
  - $Q2$ = II (*środkowy*) kwartyl = 50-ty percentyl = **mediana**
  - $Q3$ = III (*górny*) kwartyl = 75-ty percentyl

Średnia a mediana

- Średnia to **punkt koncentracji masy** zbioru danych → **wrażliwa** na obserwacje odstające
- Mediana to **wartość środkowa** zbioru danych → **nie jest wrażliwa** na obserwacje odstające



Źródło: `https://300gospodarka.pl/opinie/przecietne-wynagrodzenie-gus-interpretacja`

Rozstęp i odstęp międzykwartylowy

- Rozstęp

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

- Odstęp międzykwartylowy ($IQR =$ *inter-quartile range*)

$$IQR = Q3 - Q1$$

Wariancja

- Przeciętne **kwadratowe** odchylenie wyników od średniej
- Wariancja w próbie (estymator nieobciążony)

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

- Wariancja w populacji

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Odchylenie standardowe

- Sprawdzenie wariancji do *podstawowej jednostki* $\rightarrow$ łatwiejsza interpretacja
- Odchylenie standardowe w próbie

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

- Odchylenie standardowe w populacji

$$\sigma = \sqrt{\sigma^2}$$

Współczynnik zmienności

- Odchylenie standardowe to miara **absolutna** $\rightarrow$ problem przy porównywaniu różnych zbiorów danych
- Współczynnik zmienności (CV = *coefficient of variation*) to miara **względna**

$$CV = \frac{s}{\bar{x}}$$

Moment centralny

- Moment centalny rzędu k

$$M_k = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^k}{n}$$

- M_2 to wariancja (estymator obciążony)

Skośność

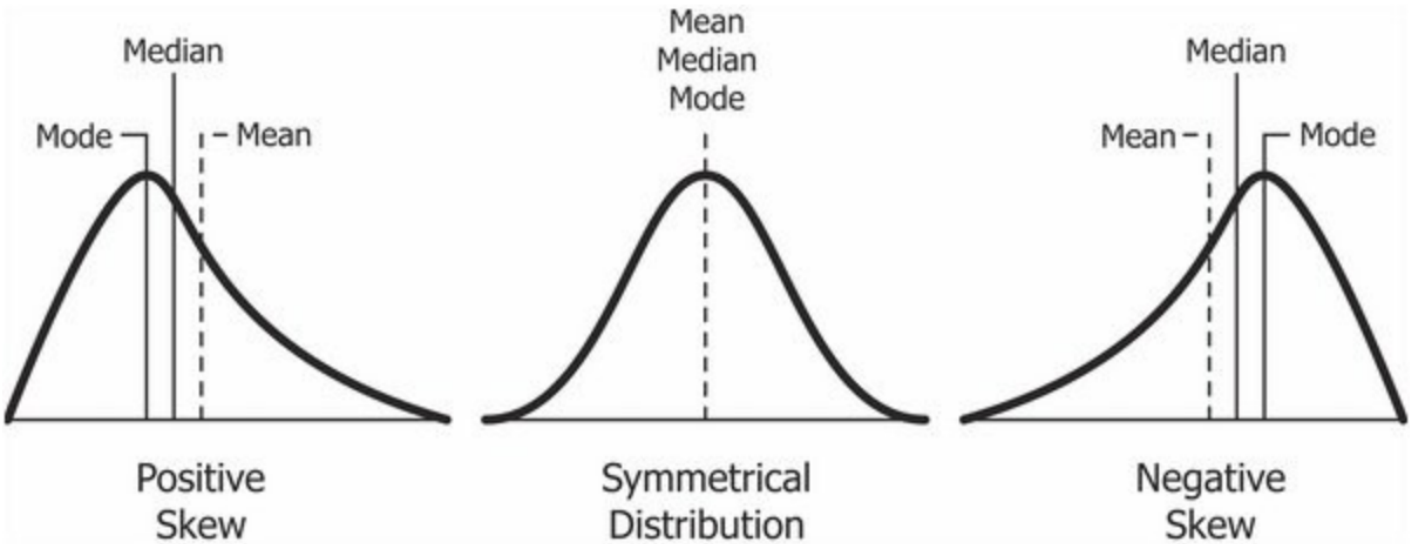
Skośność (*skewness*) to miara *asymetrii* rozkładu częstości

$$skew = \frac{M_3}{s^3} \approx \frac{M_3}{M_2^{3/2}} \approx \frac{\bar{x} - d}{s} \approx 3\frac{\bar{x} - m}{s}$$

gdzie d – dominanta, m – mediana

- rozkład symetryczny $\rightarrow skew = 0$, $\bar{x} = m = d$
- rozkład *prawoskośny* (skośność prawostronna) $\rightarrow skew > 0$, $d < m < \bar{x}$
- rozkład *lewoskośny* (skośność lewostronna) $\rightarrow skew < 0$, $\bar{x} < m < d$

Skośność



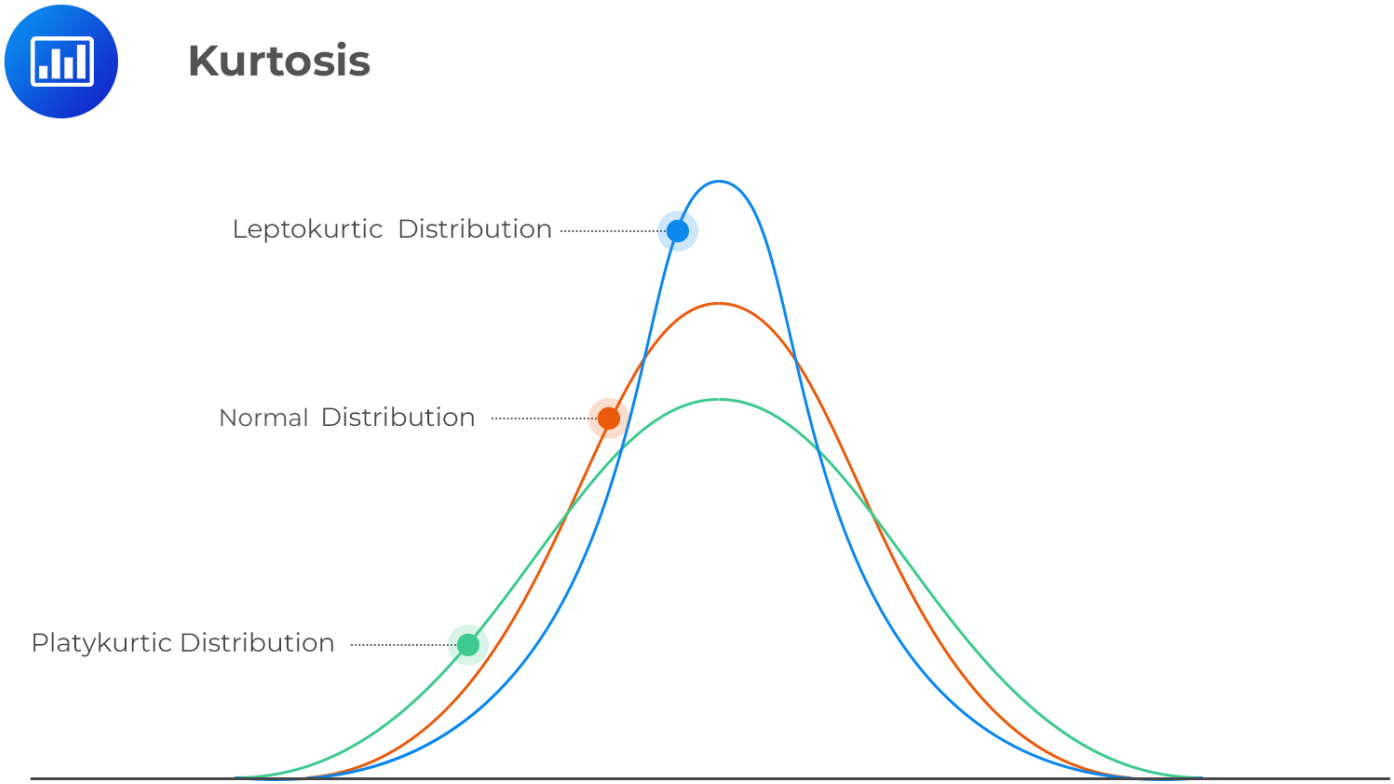
Źródło: `https://en.wikipedia.org/wiki/Skewness`

Kurtoza

Kurtoza (*kurtosis*) to miara spłaszczenia rozkładu częstości
$kurt = \frac{M_4}{s^4} - 3 \approx \frac{M_4}{M_2^2} - 3$

- rozkład *mezokurtyczny* ("normalny") $\rightarrow kurt = 0$
- rozkład *platokurtyczny* ("spłaszczony") $\rightarrow kurt < 0$
- rozkład *leptokurtyczny* ("spiczasty") $\rightarrow kurt > 0$

Kurtoza



Źródło: https://analystprep.com/cfa-level-1-exam/quantitative-methods/kurtosis-and-skewness-types-of-distributions/

Średnia i wariancja w szeregu przedziałowym

- Przyjęcie środka i-tego przedziału klasowego $\dot{x}_i$ jako reprezentata tego przedziału, który występuje n_i razy
- Średnia

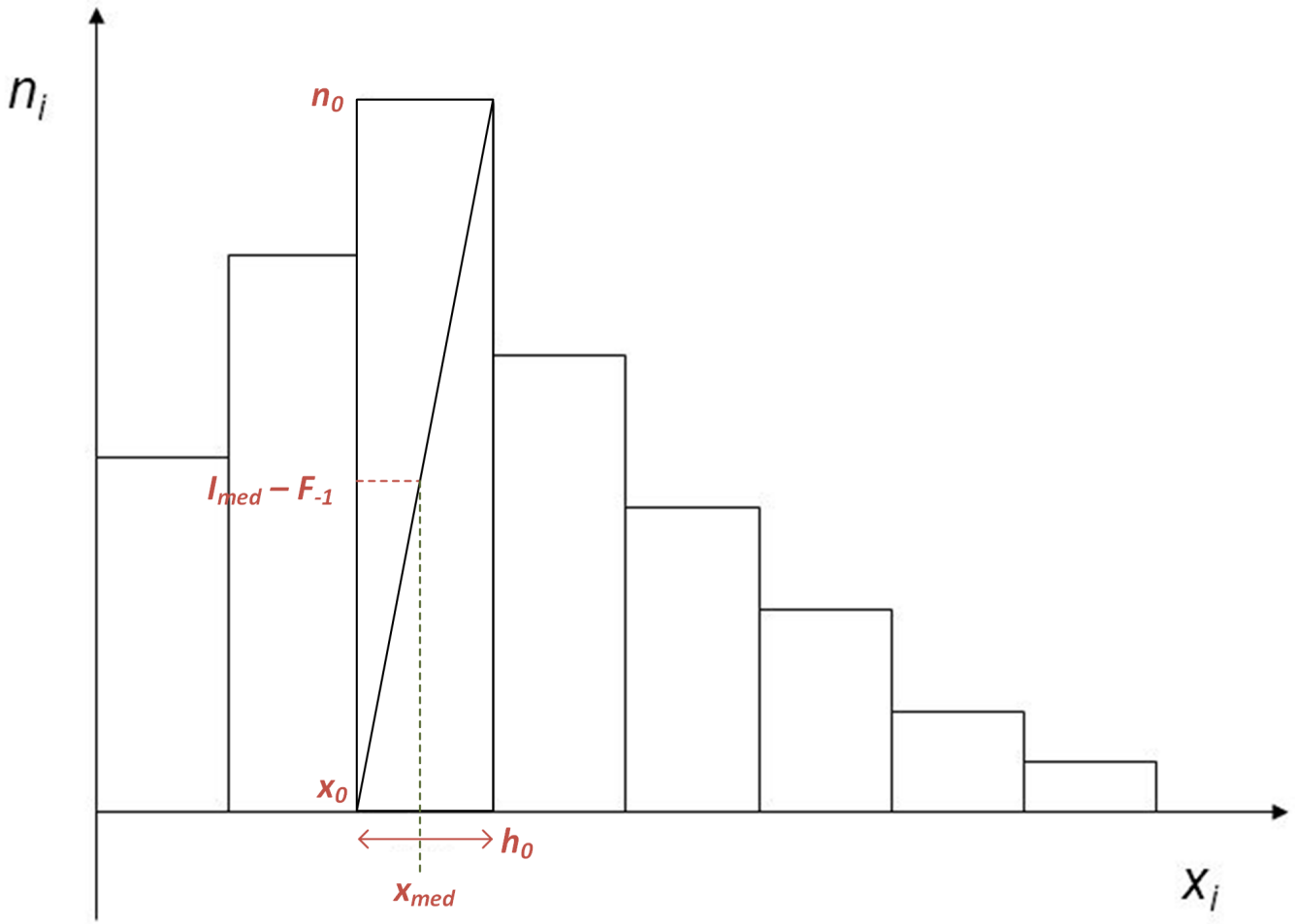
$$\bar{x} = \frac{\sum_{i=1}^{k} n_i \dot{x}_i}{n} = \frac{\sum_{i=1}^{k} n_i \dot{x}_i}{\sum_{i=1}^{k} n_i}$$

- Wariancja

$$s^2 = \frac{\sum_{i=1}^{k} n_i \left(\dot{x}_i - \bar{x}\right)^2}{n-1}$$

Mediana w szeregu przedziałowym

$$x_{med} = x_0 + \frac{h_0}{n_0} \times (i_{med} - F_{-1})$$



Dominanta w szeregu przedziałowym

$$x_{dom} = x_0 + h_0 \times \frac{n_0 - n_{-1}}{n_0 - n_{-1} + n_0 - n_{+1}}$$