

POLITECHNIKA WROCŁAWSKA
WYDZIAŁ INFORMATYKI I TELEKOMUNIKACJI

METODY ANALIZY I EKSPLORACJI DANYCH

Wykład 3 - Statystyki opisowe

DR INŻ. AGATA MIGALSKA



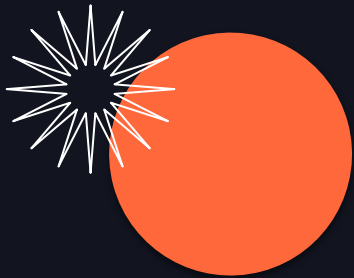
Wykład
3

01
**MIARY
POŁOŻENIA**

02
**MIARY
ZRÓŻNICOWANIA**

03
**MIARY
KORELACJI**

04
**KURTOZA I
SKOŚNOŚĆ**



CEL I MOTYWACJA

STATYSTYKA OPISOWA

- Zajmuje się metodami opisu danych statystycznych uzyskanych podczas badania statystycznego.
 - Pozwala na podsumowanie zbioru danych i wyciągnięcie podstawowych wniosków i uogólnień na temat zbioru.
-

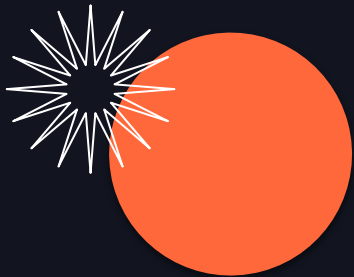
STATYSTYKA OPISOWA

Jedna zmienna

- Obejmuje opisanie rozkładu pojedynczej zmiennej: jej tendencję centralną i rozproszenie.
- Kształt rozkładu można również opisać za pomocą wskaźników, takich jak skośność i kurtoza.
- Charakterystykę rozkładu zmiennej można również przedstawić w formie graficznej lub tabelarycznej, jako histogramu lub wykresu "łodyga z liśćmi".

Wiele zmiennych

- Opis poszczególnych zmiennych (patrz "statystyka opisowa jednej zmiennej")
 - Graficzna reprezentacja za pomocą wykresów punktowych dla każdej pary zmiennych
 - Współczynniki korelacji pomiędzy zmiennymi
-



01

MIARY POŁOŻENIA

MIARY POŁOŻENIA

TENDENCJA CENTRALNA

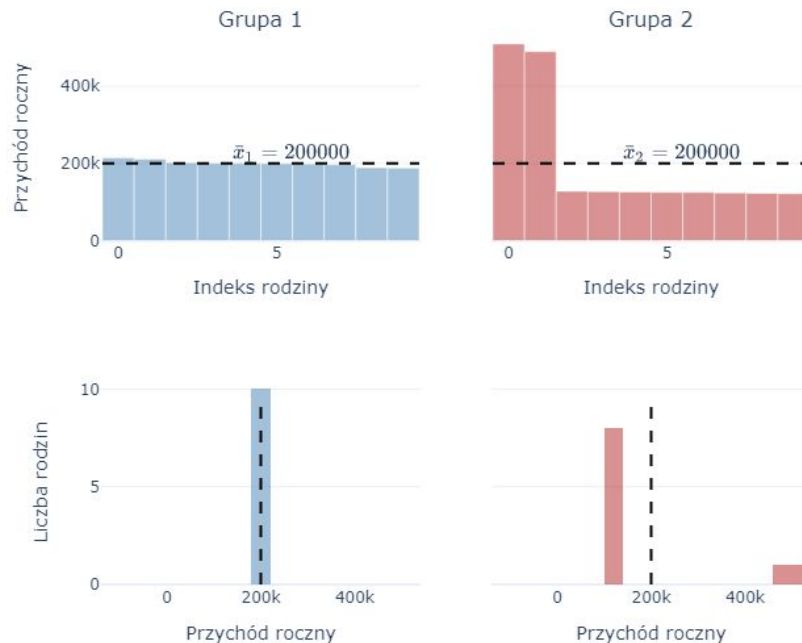


ŚREDNIA ARYTMETYCZNA

- Zbiór liczb x_1, x_2, \dots, x_N
- N - liczność zbioru
- Średnia arytmetyczna

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

Roczny przychód na rodzinę





ŚREDNIA GEOMETRYCZNA

- Zbiór liczb x_1, x_2, \dots, x_N
- Wykorzystywana przede wszystkim wtedy, gdy ma się do czynienia z wielkościami zmieniającymi się w postępie geometrycznym, tzn. gdy kolejna wielkość w szeregu powstaje przez pomnożenie przez stały mnożnik wielkości bezpośrednio ją poprzedzającej.
 - gdy wartości zmiennej są wartościami względnymi (wskaźniki, współczynniki, etc)
- Znaczenie w:
 - ekonomii i inwestowaniu,
 - demografii.
- Nie zachowuje jednostek (patrz przykład w ocenami na następnym slajdzie)

$$\overline{x}_G = \sqrt[N]{\prod_{i=1}^N x_i}$$

$$\log \overline{x}_G = \frac{1}{N} \sum_{i=1}^N \log x_i$$



ŚREDNIA GEOMETRYCZNA

- **Przykład:** współczynnik inflacji obliczany przez GUS
- **Przykład:** Rozważmy portfel akcji, który wzrasta ze \$100 do \$110 w pierwszym roku, następnie spada do \$80 w drugim i wzrasta do \$150 w trzecim roku. Roczny zwrot z portfela wynosi:
 $((110/100) * (80/110) * (150/110))^{1/3} = 0,1447 = 14,47\%$.
- **Przykład:** Chcemy porównać oceny dwóch kawiarni z dwóch różnych źródeł. Problem polega na tym, że źródło 1 używa 5-gwiazdkowej skali, a źródło 2 używa 100-punktowej skali. Która kawiarnia jest lepiej oceniana?

	Źródło 1	Źródło 2
Kawiarnia 1	4.5	68
Kawiarnia 2	3	75



ŚREDNIA GEOMETRYCZNA

- **Przykład:** współczynnik inflacji obliczany przez GUS
- **Przykład:** Rozważmy portfel akcji, który wzrasta ze \$100 do \$110 w pierwszym roku, następnie spada do \$80 w drugim i wzrasta do \$150 w trzecim roku. Roczny zwrot z portfela wynosi:
 $((110/100) \cdot (80/110) \cdot (150/110))^{1/3} = 0,1447 = 14,47\%$.
- **Przykład:** Chcemy porównać oceny dwóch kawiarni z dwóch różnych źródeł. Problem polega na tym, że źródło 1 używa 5-gwiazdkowej skali, a źródło 2 używa 100-punktowej skali. Która kawiarnia jest lepiej oceniana?

	Źródło 1	Źródło 2	Średnia geometryczna
Kawiarnia 1	4.5	68	$\sqrt{4.5 \cdot 68} = 17.5$
Kawiarnia 2	3	75	$\sqrt{3 \cdot 75} = 15$



ŚREDNIA HARMONICZNA

- Zbiór liczb x_1, x_2, \dots, x_N
- Wykorzystywana przede wszystkim wtedy, gdy dane przedstawione są w postaci względnej:
 - prędkość przedstawiana w km/h lub m/s
 - prędochłonność w min/sztukę
 - spożycie w kg/osobę lub litrach/osobę
 - cena jednostkowa w zł/szt
 - czyli wartości cechy przedstawiamy w przeliczeniu na stałą jednostkę innej zmiennej.

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



ŚREDNIA HARMONICZNA

- **Przykład:** Rowerzysta jedzie do pracy godzinę z prędkością 20 km/h. Po pracy, zmęczony, wraca 2 godziny, ze średnią prędkością 10km/h. Jaka jest średnia prędkość rowerzysty?
 - Średnia arytmetyczna wynosi 15 km/h ale to by znaczyło, że na przejechanie 40km rowerzysta potrzebowałby 2h40, a wiemy, że jechał 3h.
 - Średnia harmoniczna wynosi $\frac{2}{\frac{1}{20} + \frac{1}{10}} = \frac{40}{3} = 13.33$ km/h, co daje 3h na przejechanie 40km, czyli dokładnie tyle ile potrzebował nasz rowerzysta.
- **Przykład:** Dwa rezystory o oporze 40Ω i 60Ω połączone równolegle można zastąpić dwoma rezystorami o oporze $\frac{2}{\frac{1}{60} + \frac{1}{40}} = \frac{2 \cdot 120}{5} = 48$ Ω każdy.

MEDIANA

- Zbiór liczb x_1, x_2, \dots, x_N uporządkowany od najmniejszej do największej
- N - liczność zbioru
- Mediana

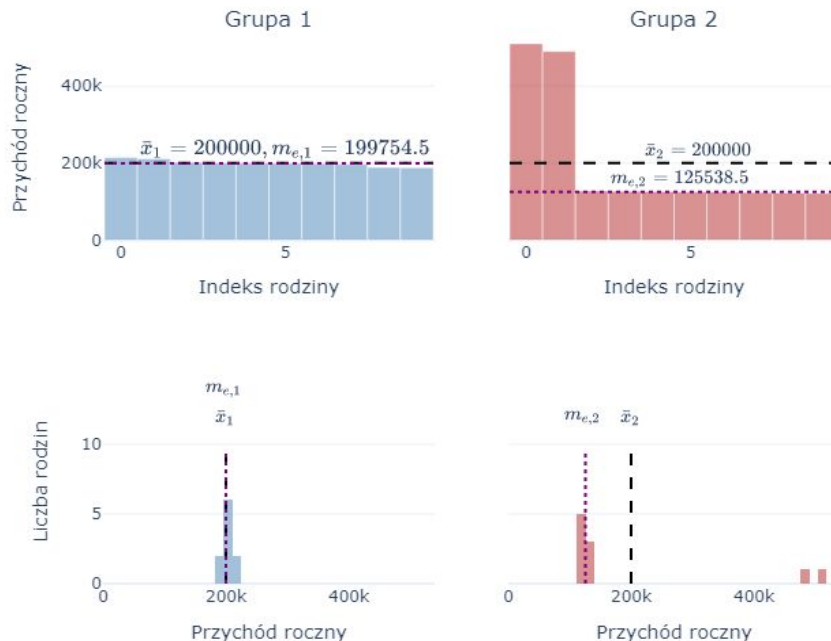
- Gdy N jest liczbą nieparzystą

$$m_e = x_{\frac{N+1}{2}}$$

- Gdy N jest liczbą parzystą,

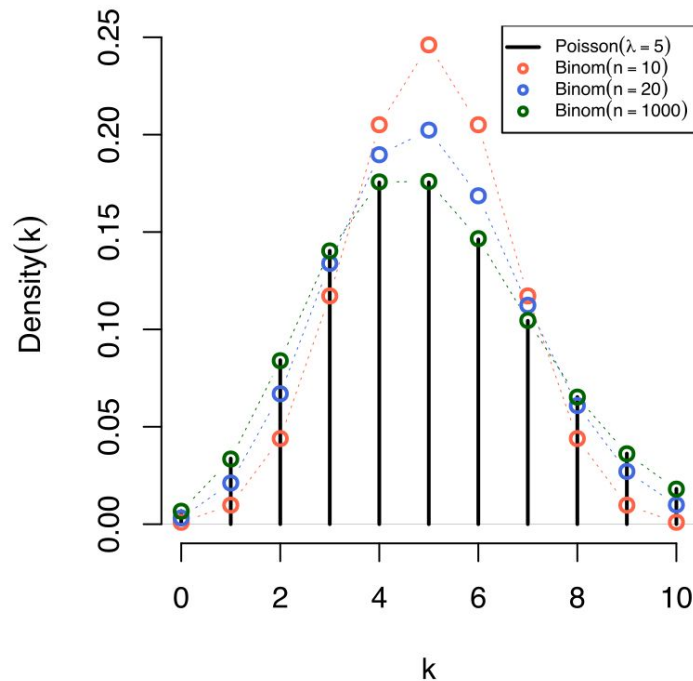
$$m_e = \frac{x_{N/2} + x_{(N+1)/2}}{2}$$

Roczny przychód na rodzinę



MODA

- Inne nazwy: dominanta, wartość modalna, wartość najczęstsza
- Statystyka dla zmiennych o wartościach dyskretnych
- Modą nazywamy wartość najczęściej występującą w zbiorze danych tzn. wartość x , dla której funkcja masy prawdopodobieństwa osiąga wartość najwyższą



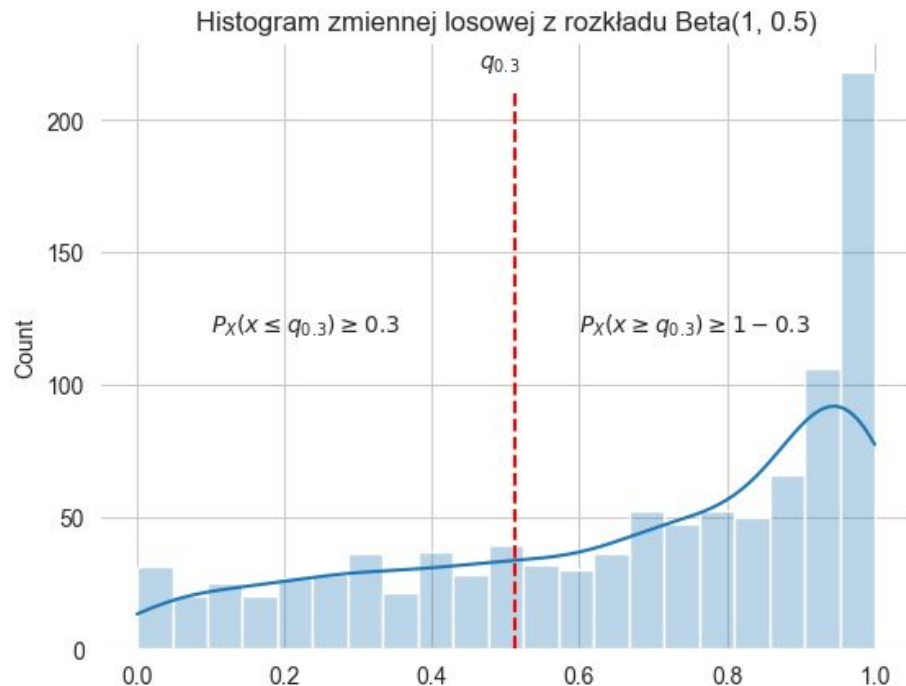
KWANTYLE

Kwantylem rzędu p ,
gdzie $0 \leq p \leq 1$, w rozkładzie
empirycznym P_X zmiennej
losowej X nazywamy każdą
liczbę x_p , dla której spełnione
są nierówności

$$P_X((-\infty, q_p]) \geq p$$

oraz

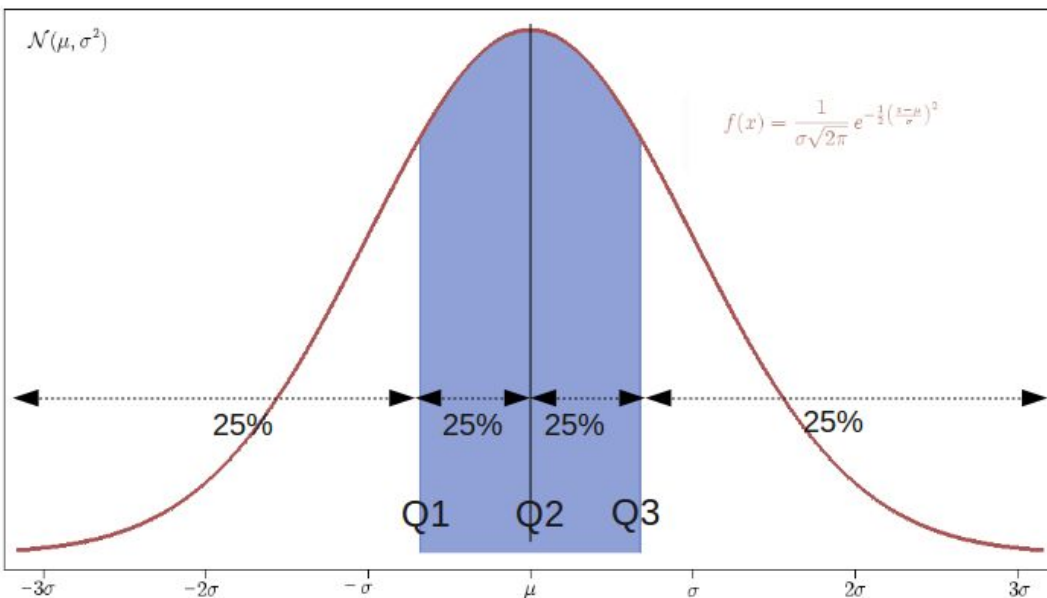
$$P_X([q_p, \infty)) \geq 1 - p$$

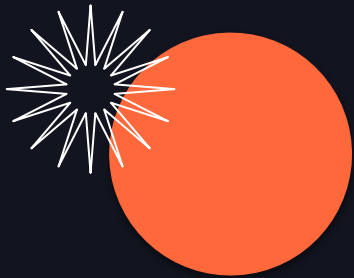


KWARTYLE

Kwantyle dzielące obserwacje na (mniej więcej) równe ćwiartki (kwarty):

- Q1 - kwantyl rzędu $\frac{1}{4}$ (0.25)
= dolny kwartyl
- Q2 - kwantyl rzędu $\frac{1}{2}$ (0.5)
= mediana
- Q3 - kwantyl rzędu $\frac{3}{4}$ (0.75)
= górny kwartyl





02

MIARY ZRÓŻNICOWANIA



ROZSTĘP

$$R = X_{\max} - X_{\min}$$

Przykład: W Irkucku amplituda średnich temperatur powietrza wynosi $38,4^{\circ}\text{C}$ - średnia miesięczna temperatura powietrza jest najwyższa w lipcu i wynosi $+17,5^{\circ}\text{C}$, a najniższa temperatura w styczniu wynosi $-20,9^{\circ}\text{C}$.

$$R = X_{\max} - X_{\min} = 17.5^{\circ}\text{C} - (-20.9^{\circ}\text{C}) = 38.4^{\circ}\text{C}$$

Przykład: „Ceny tej książki wahają się od 20 do 25 złotych. Oznacza to różnicę aż 5 zł na sztuce.”
(wartość minimalna, wartość maksymalna, rozstęp)

Przykład: Widełki płacowe

ODCHYLENIE STANDARDOWE

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

- **Odchylenie standardowe** mówi o tym, o ile średnio odchylają się wartości badanej cechy od średniej arytmetycznej.
- Odchylenie standardowe liczymy, żeby stwierdzić, czy w naszej populacji jednostki są podobne ze względu na badaną cechę, czy znacznie różnią się między sobą.





WARIANCJA

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

- Wartość wariancji jest wyrażona w kwadracie jednostki
 - wariancję średniego wzrostu mamy w metrach kwadratowych,
 - wariancja średniego zużycia długopisów podana jest w... długopisach kwadratowych.
 - Dużo wygodniej jest wyciągnąć z tej wartości pierwiastek i operować wskaźnikiem, który ma taką samą jednostkę jak analizowana cecha (czyli odchyleniem standardowym).
-



ODCHYLENIE BEZWZGLĘDNE

- Miary zróżnicowania odporne na obserwacje odstające
- Ten sam symbol (MAD) często stosowany do obu miar, powodując konfuzję
- **Mediana odchylenia bezwzględnego wokół mediany**

$$MAD_e = m_e(|x_i - m_e(x)|)$$

- miara sugerowana we współczesnej literaturze do wyznaczania obserwacji odstających
- teoretycznie może być wokół średniej, ale rzadko stosowana
- **Średnie odchylenie bezwzględne wokół średniej / mediany**

$$MAD = \frac{\sum_{i=1}^N |x_i - m(x)|}{N}$$

- “klasyczna” wersja
- $m(x)$ to średnia arytmetyczna lub mediana.

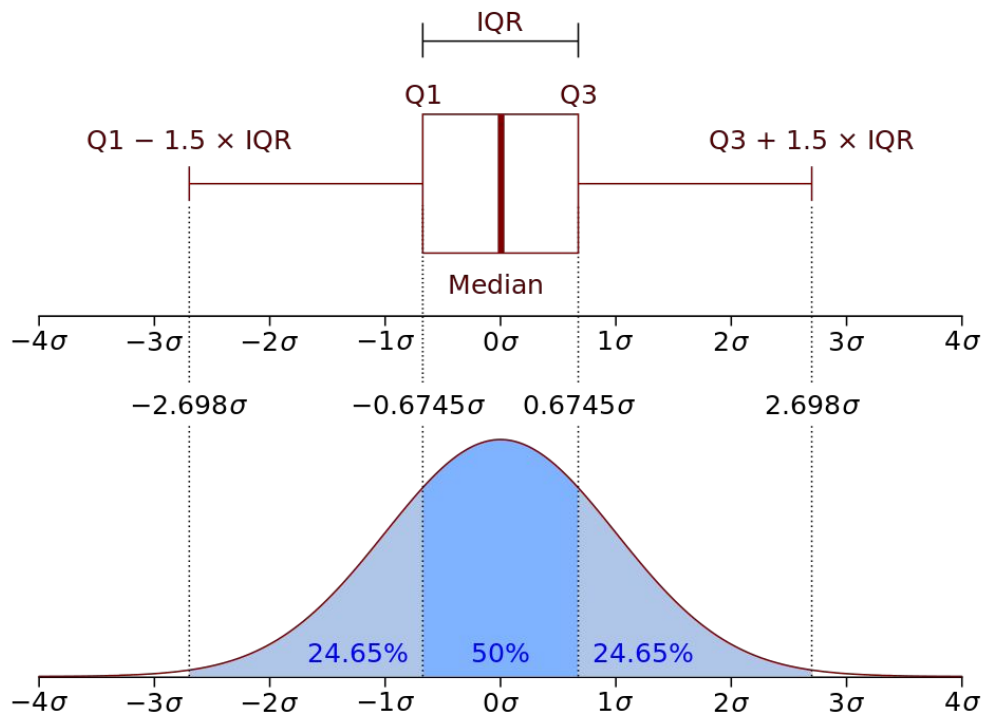


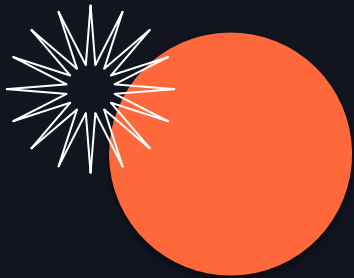
ROZSTĘP KWARTYLNÝ = ĆWIARTKOWY

$$IQR = Q_3 - Q_1$$

- Jest to miara, która sporo mówi o populacji, gdyż w tych granicach mieści się 50% badanych obiektów.
 - Im większy rozstęp kwartylny, tym bardziej zróżnicowana jest cecha statystyczna.
-

POWTÓRKA: WYKRES PUDEŁKOWY





03

MIARY KORELACJI



KORELACJA

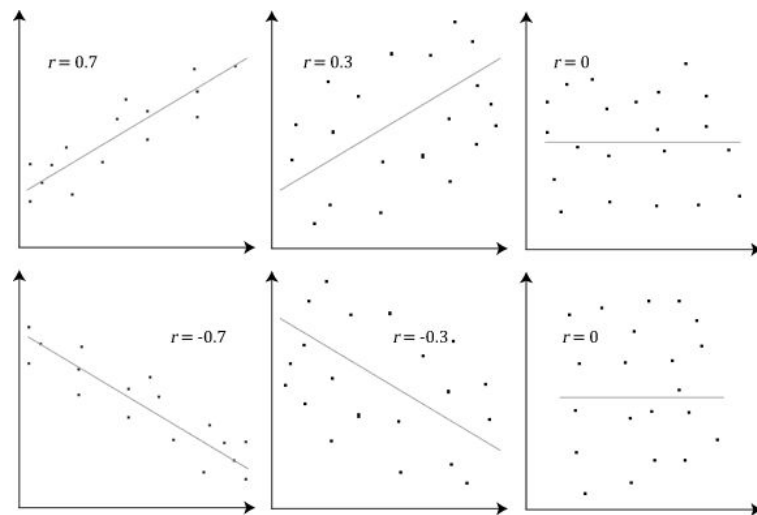
- Mierzy siłę powiązania między dwiema zmiennymi i kierunek związku.
 - Wartość współczynnika korelacji waha się od +1 do -1.
 - Wartość ± 1 wskazuje na doskonały stopień powiązania między tymi dwiema zmiennymi.
 - Wraz ze zbliżaniem się wartości współczynnika korelacji do zera, związek między dwiema zmiennymi będzie słabszy.
 - Kierunek zależności wskazuje znak współczynnika;
 - znak + oznacza pozytywną relację,
 - znak – negatywną relację.
-

WSPÓŁCZYNNIK KORELACJI PEARSONA

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}$$

Założenia:

- Każda obserwacja powinna mieć parę wartości.
- Zmienne są ciągłe.
- Brak wartości odstających.
- Zakłada liniowość i homoskedastyczność (wszystkie zmienne losowe w tym ciągu posiadają tę samą, skończoną wariancję).



WSPÓŁCZYNNIK KORELACJI SPEARMANA

$$\rho = \frac{\sum_{i=1}^N (R(x_i) - \overline{R(x)}) (R(y_i) - \overline{R(y)})}{\sqrt{\sum_{i=1}^N (R(x_i) - \overline{R(x)})^2 \sum_{i=1}^N (R(y_i) - \overline{R(y)})^2}}$$
$$= 1 - \frac{6 \sum_{i=1}^N (R(x_i) - R(y_i))^2}{N(N^2 - 1)}$$

Założenia:

- Pary obserwacji są niezależne.
- Dwie zmienne powinny być mierzone na skali porządkowej, interwałowej lub ilorazowej.
- Zakłada, że istnieje monotoniczna zależność między tymi dwiema zmiennymi.

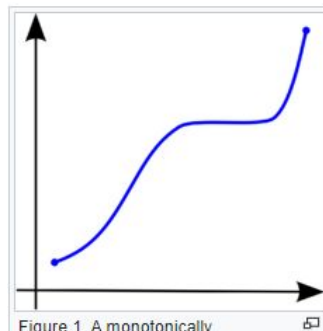


Figure 1. A monotonically increasing function.

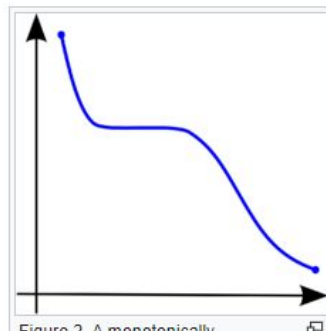


Figure 2. A monotonically decreasing function

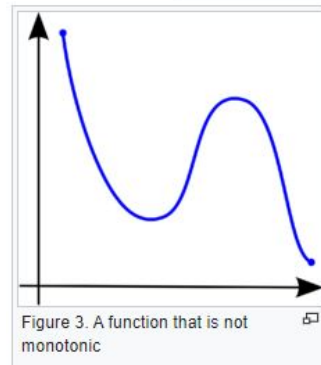


Figure 3. A function that is not monotonic



WSPÓŁCZYNNIK KORELACJI KENDALLA

$$\tau = \frac{n_c - n_d}{n_c + n_d}$$

n_c - liczba par zgodnych,

n_d - liczba par niezgodnych.

Założenia (takie same jak dla Spearmana):

- Pary obserwacji są niezależne.
 - Dwie zmienne powinny być mierzone na skali porządkowej, interwałowej lub ilorazowej.
 - Zakłada, że istnieje monotoniczna zależność między tymi dwiema zmiennymi.
-



KTÓRY WSPÓŁCZYNNIK WYBRAĆ?

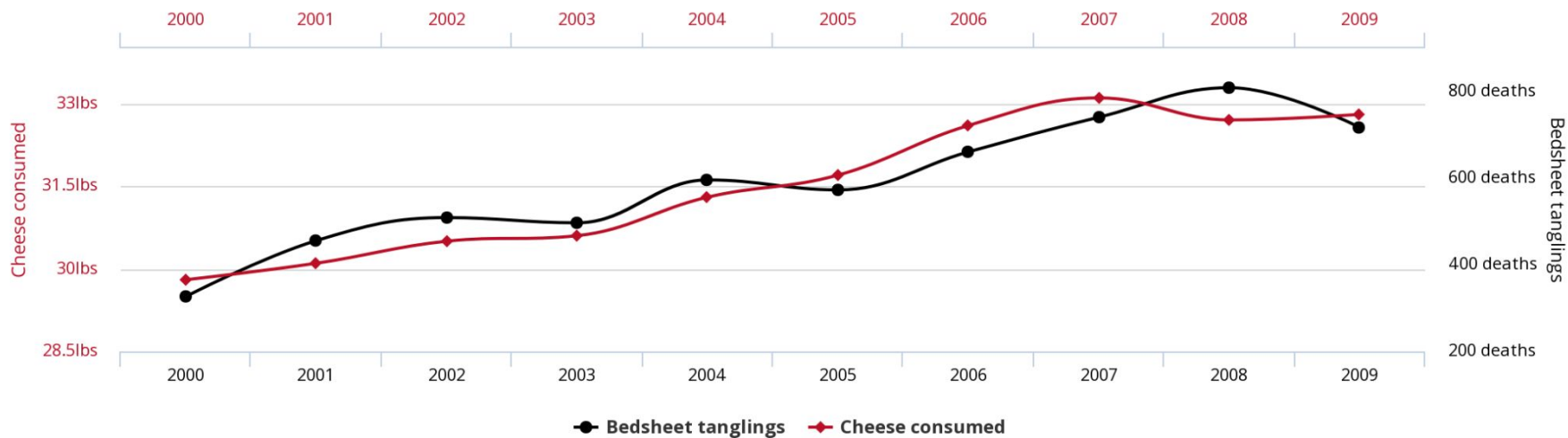
- Pearson vs Spearman vs Kendall
 - Korelacje nieparametryczne mają mniejszą moc, ponieważ w swoich obliczeniach wykorzystują mniej informacji. W przypadku korelacji Pearsona wykorzystuje się informacje o średniej i odchyleniu od średniej, natomiast korelacje nieparametryczne wykorzystują jedynie informacje porządkowe i wyniki par.
 - wybieramy najmocniejszy możliwy współczynnik w danej sytuacji
 - Współczynniki korelacji mierzą tylko relacje liniowe (Pearson) lub monotoniczne (Spearman i Kendall).
 - Spearman vs Kendall
 - Współczynnik korelacji Kendalla jest bardziej odporny na małe licznosci próbek oraz na obserwacje odstające niż współczynnik korelacji Spearmana.
-

FAŁSZYWE KORELACJE

Per capita cheese consumption
correlates with

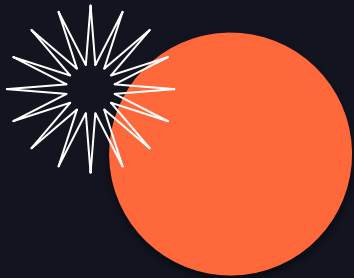
Korelacja: 94.71%

Number of people who died by becoming tangled in their bedsheets



tylervigen.com

<http://www.tylervigen.com/spurious-correlations>

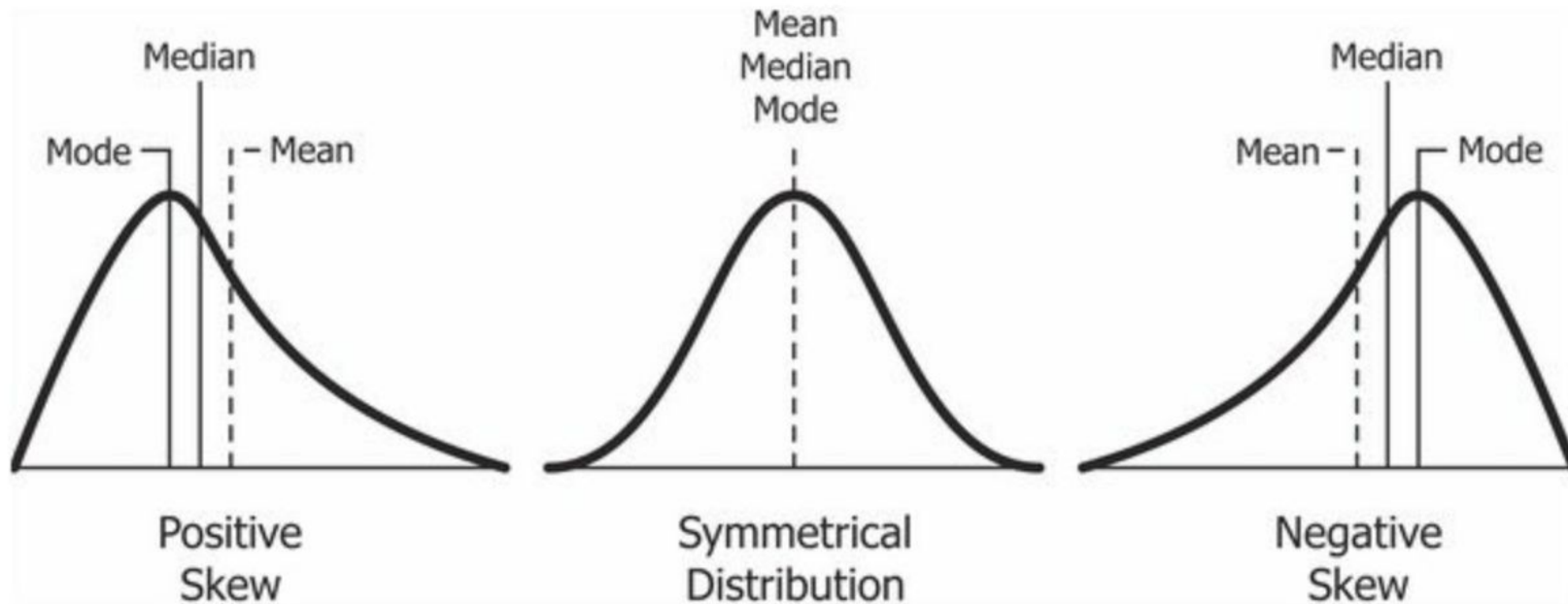


03

KURTOZA I SKOŚNOŚĆ



SKOŚNOŚĆ



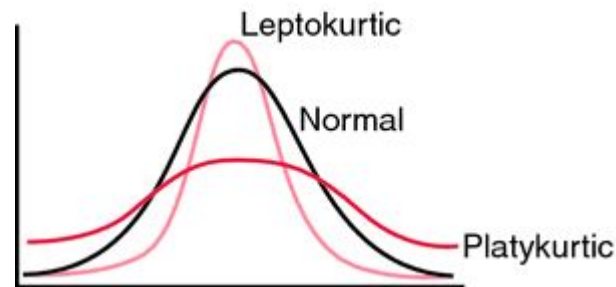
KURTOZA



ang. platykurtic platypus



ang. leptokurtic lepping kangaroos



- **mezokurtyczne** ($K = 0$) – intensywność wartości skrajnych jest podobna do intensywności wartości skrajnych rozkładu normalnego (dla którego kurtoza wynosi dokładnie 0)
- **leptokurtyczne** ($K > 0$) – intensywność wartości skrajnych jest większa niż dla rozkładu normalnego („ogony” rozkładu są „grubsze”)
- **platykurtyczne** ($K < 0$) – intensywność wartości skrajnych jest mniejsza niż w przypadku rozkładu normalnego („ogony” rozkładu są „węższe”).

THANKS!

**DZIĘKUJĘ
ZA UWAGĘ**

