

Wektor



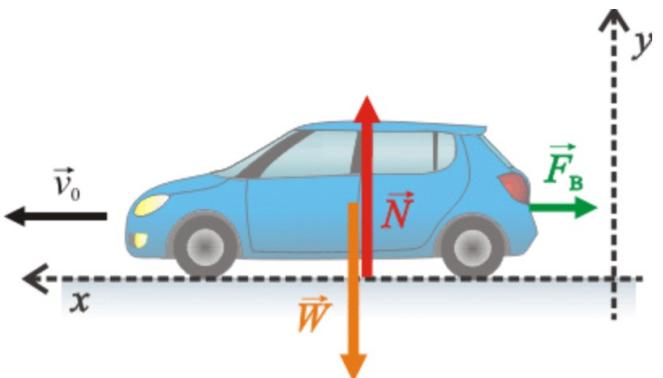
Programista

Tablica wypełniona liczbami

`array([1, 2, 3])`

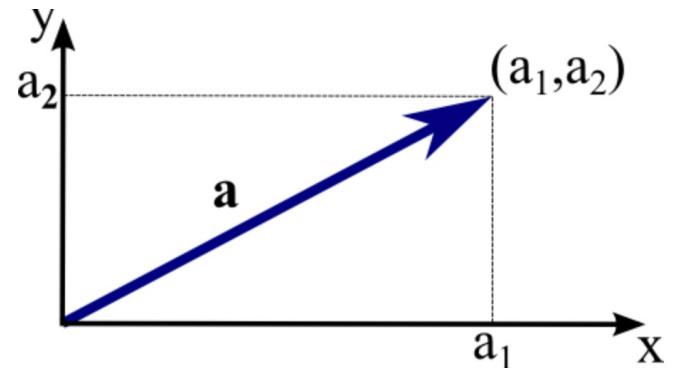
Fizyk

Obiekt opisywany za pomocą
długości, kierunku i zwrotu.



Matematyk

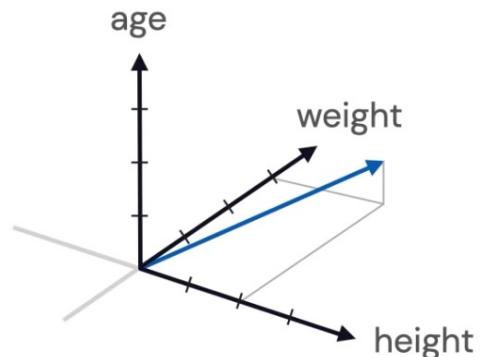
Współrzędne punktu w przestrzeni.



Wektor – reprezentacja danych

$$\mathbf{p} = \begin{bmatrix} 64 \\ 131 \\ 23 \end{bmatrix} \begin{array}{l} \text{height} \\ \text{weight} \\ \text{age} \end{array}$$

“p” for “patient”



A 5x5 grid representing a convolutional layer's receptive field. The input values are shown in light gray, while the output values (the result of applying a 3x3 kernel) are shown in black. The output values are placed at indices (1,1), (1,3), (3,1), (3,3), (4,1), and (4,3). This illustrates a stride of 2.

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \textcolor{red}{1} \\ 0 \\ 0 \\ \textcolor{red}{1} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$cat = \begin{bmatrix} 0 \\ 0 \\ 2 \\ 0 \\ 15 \\ 1 \\ 0 \\ 0 \\ 51 \\ \vdots \\ 0 \end{bmatrix} \quad \begin{array}{l} \text{doc \#1} \\ \text{doc \#2} \\ \text{doc \#3} \\ \text{doc \#4} \\ \text{doc \#5} \\ \text{doc \#6} \\ \text{doc \#7} \\ \text{doc \#8} \\ \text{doc \#9} \\ \dots \\ \text{doc \#1500} \end{array}$$

Tensor - przykład



8,11,0, 55,13,25,19

15,241,2,155,13,35,65

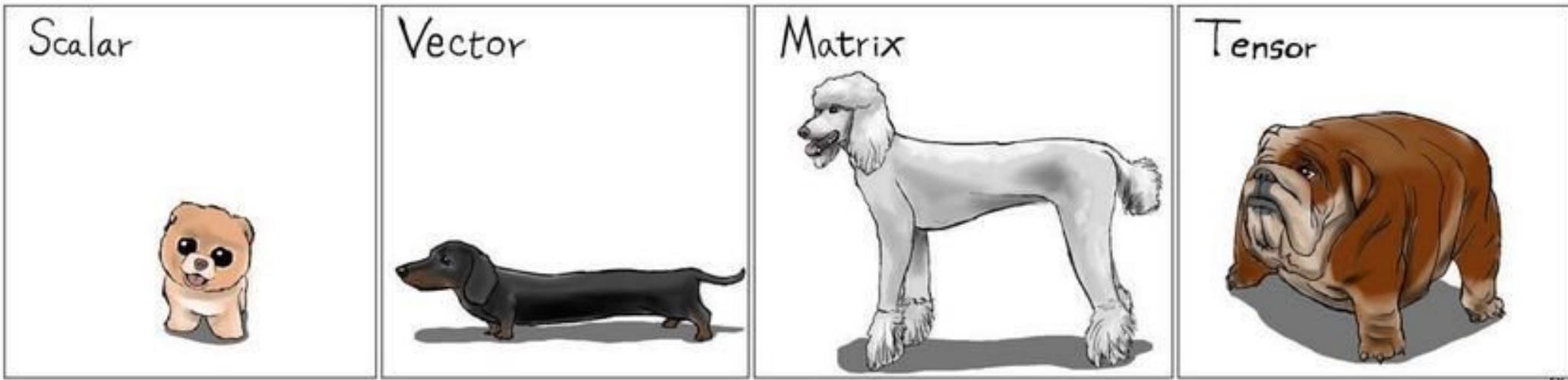
14,211,0,255,23,45,11

05,255,1,255,10,17,23

77,167,9,112,56,16,90

45,245,0,145,22,55,48

Tensor - reprezentacja



9

array([9, -8, 5])

array([[4, 7, -4],
 [-8, 4, 9],
 [7, 6, 6]])

array([[[7, 0, -10],
 [7, -2, 7],
 [-3, -1, -8]],

 [[-2, 8, -9],
 [6, 6, -8],
 [-9, -6, -6]],

 [[-6, -7, -10],
 [-2, -6, 6],
 [0, -4, 4]]])

Mnożenie macierzy

$$X = \begin{bmatrix} 5 & 6 & 1 & 2 \\ 8 & 7 & 6 & 3 \\ 5 & 0 & 6 & 4 \end{bmatrix}$$

$$Y = \begin{bmatrix} 3 & 0 & 4 & 9 \\ 4 & 6 & 5 & 8 \\ 7 & 0 & 1 & 5 \end{bmatrix}$$

$$Z = X * Y = \begin{bmatrix} 15 & 0 & 4 & 18 \\ 32 & 42 & 30 & 24 \\ 35 & 0 & 6 & 20 \end{bmatrix}$$

Każdy element macierzy X jest pomnożony przez odpowiadający mu element macierzy Y

$$\begin{array}{lll} x_{11} = 5 & y_{11} = 3 & z_{11} = x_{11} \times y_{11} = 15 \\ x_{12} = 6 & y_{12} = 0 & z_{12} = x_{12} \times y_{12} = 0 \end{array}$$

$$X_{3 \times 4}$$

$$Y_{3 \times 4}$$

↔ Wymiar obu macierzy musi być taki sam

Mnożenie macierzy (algebraiczne) (*dot product*)

$$\mathbf{X} = \begin{bmatrix} [9 & 2 & 2] \\ [4 & 0 & 0] \\ [9 & 3 & 9] \end{bmatrix}$$

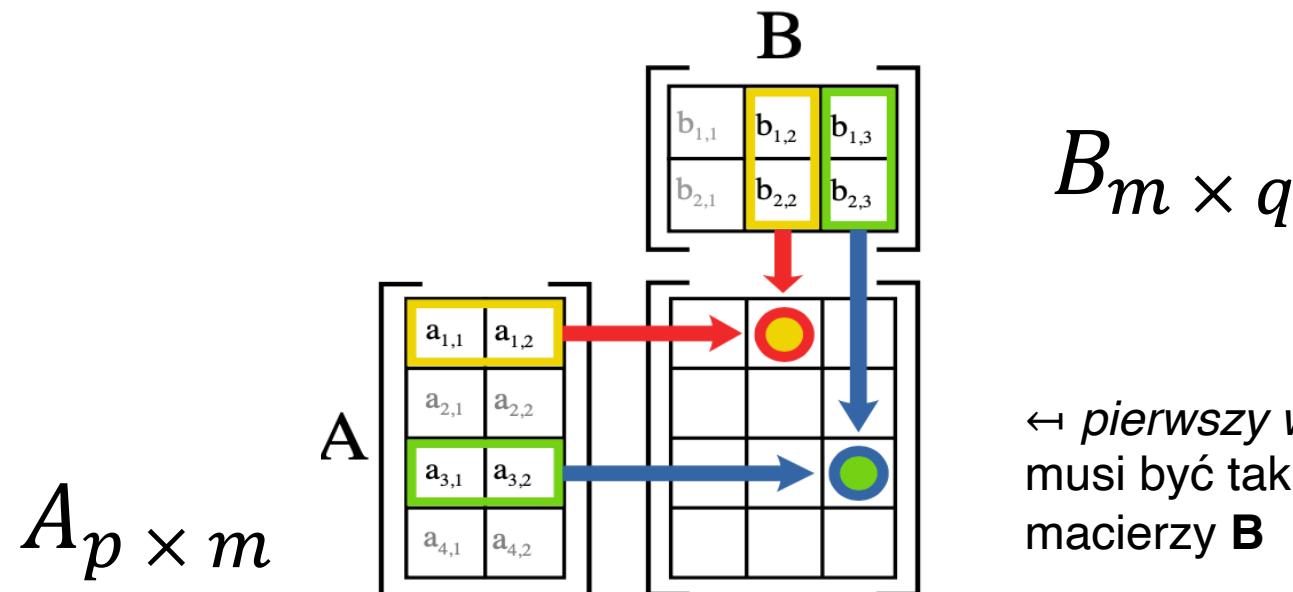
$$\mathbf{Y} = \begin{bmatrix} [8 & 1 & 1] \\ [9 & 6 & 8] \\ [7 & 4 & 8] \end{bmatrix}$$

$$\mathbf{Z} = \mathbf{X} \cdot \mathbf{Y} = \begin{bmatrix} [[104 & 29 & 41] \\ [32 & 4 & 4] \\ [162 & 63 & 105]] \end{bmatrix}$$

$$z_{11} = x_{11} \times y_{11} + x_{12} \times y_{21} + x_{13} \times y_{31}$$

$$z_{21} = x_{21} \times y_{11} + x_{22} \times y_{21} + x_{23} \times y_{31}$$

$$z_{12} = x_{11} \times y_{12} + x_{12} \times y_{22} + x_{13} \times y_{32}$$



Element neutralny i element odwrotny

* – działanie (np. dodawanie, mnożenie)

Element α nazywamy elementem neutralnym:

$$A * \alpha = \alpha * A = A$$

$$A = 5$$

* → mnożenie

$$5 \times \alpha = \alpha \times 5 = 5$$

$$\alpha = 1$$

Element neutralnym mnożenia jest 1

$$A = 3.14$$

* → dodawanie

$$3.14 + \alpha = \alpha + 3.14 = 3.14$$

$$\alpha = 0$$

Element neutralnym dodawania jest 0

Element β nazywamy elementem odwrotnym do elementu A :

$$A * \beta = \beta * A = \alpha$$

$$A = 5$$

* → mnożenie

$$5 \times \beta = \beta \times 5 = \alpha = 1$$

$$\beta = \frac{1}{5}$$

Element odwrotnym mnożenia do elementu A jest $\frac{1}{A}$

$$A = 3.14$$

* → dodawanie

$$3.14 + \beta = \beta + 3.14 = \alpha = 0$$

$$\alpha = -3.14$$

Element odwrotnym dodawania do elementu A jest $-A$

Macierz odwrotna

$A \cdot B$ – działanie mnożenia macierzy

Element neutralnym mnożenia macierzy jest macierz jednostkowa I :

$$I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

$$A \cdot I = I \cdot A = A$$

Macierz B nazywamy macierzą odwrotną do macierzy kwadratowej A gdy:

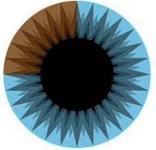
$$A \cdot B = B \cdot A = I$$

Macierz odwrotną do macierzy kwadratowej A oznaczamay:

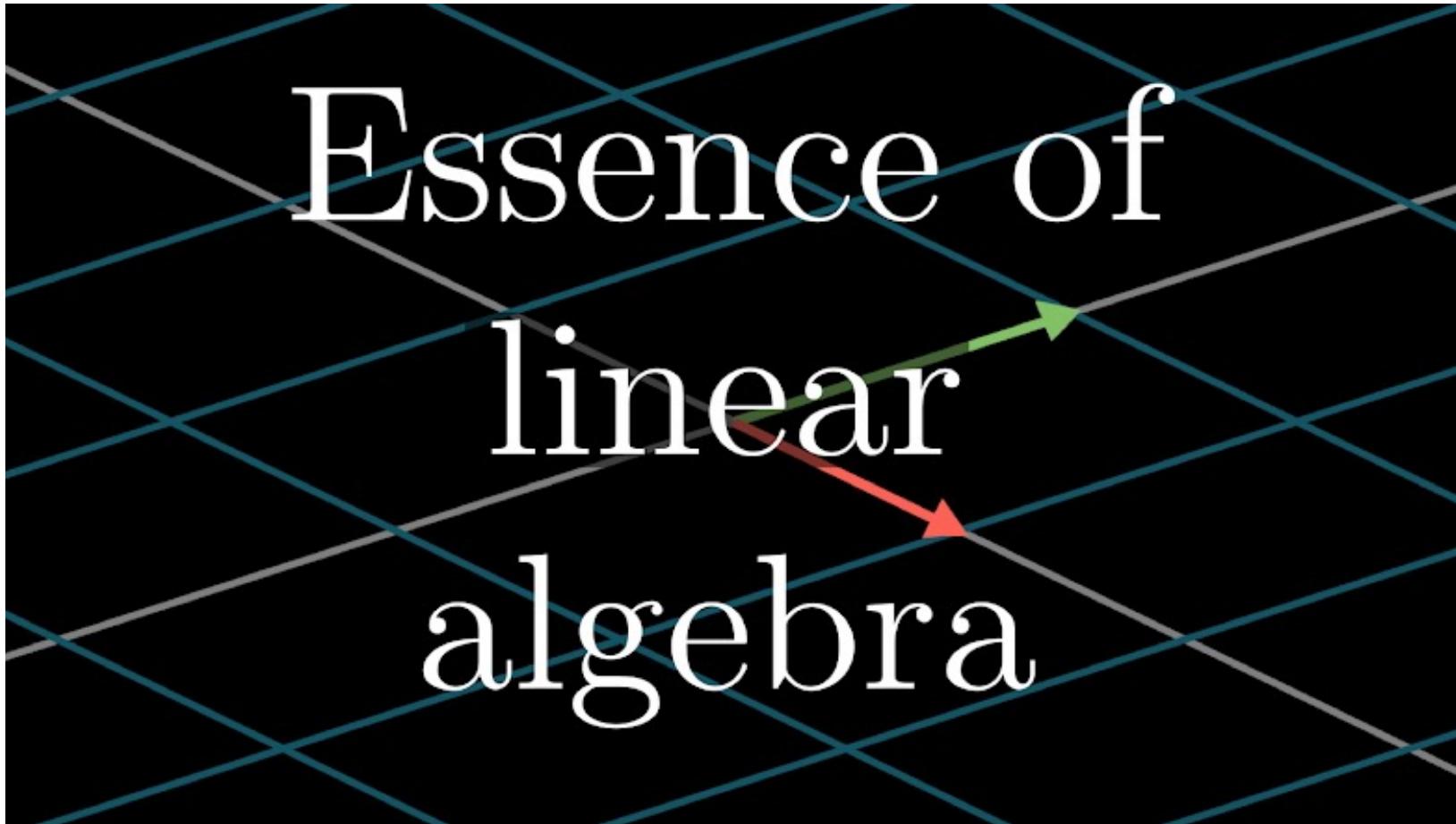
$$A^{-1}$$

Macierz odwrotną można policzyć ze wzoru:

$$A^{-1} = \frac{1}{\det(A)} (A^T)^{-1}$$



3Blue1Brown





The Matrix is everywhere. It is all around us. Even now, in this very room.

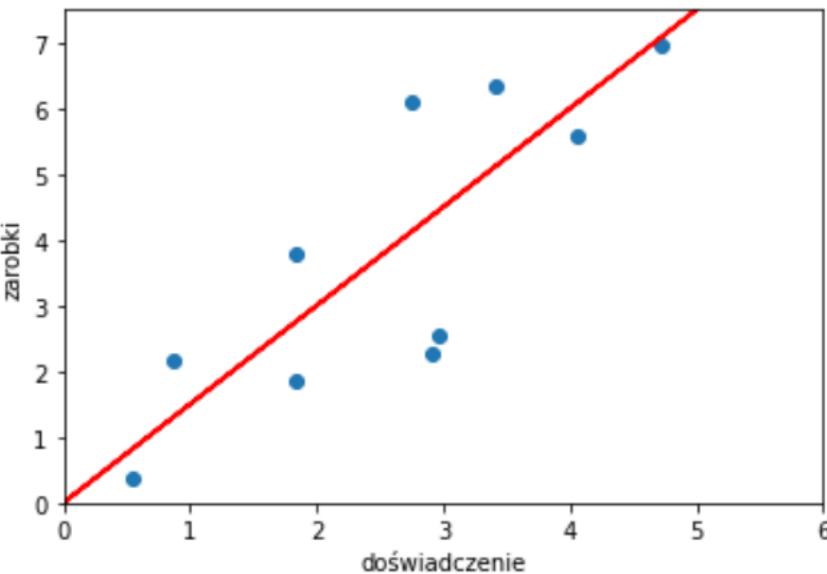
-Morpheus

Mnożenie macierzy (*geometria regresji*)

$$y = AX + b + e$$

$$\hat{y} = AX + b$$

$$\text{zarobki} = a \times \text{doświadczenie} + b$$



$$\text{zarobki}_1 = a \times \text{doświadczenie}_1 + b$$

$$\text{zarobki}_2 = a \times \text{doświadczenie}_2 + b$$

$$\text{zarobki}_3 = a \times \text{doświadczenie}_3 + b$$

(...)

b	doświadczenie	współczynniki	b	=	zarobki
1	4.057534		a_1 * 4.06 + b		5.586908
1	2.750049		a_2 * 2.75 + b		6.105262
1	2.964347		a_3 * 2.96 + b		2.533063
1	3.410334		a_4 * 3.41 + b		6.332447
1	1.828527	*		=	1.840721
1	0.548475		a_5 * 1.83 + b		0.355877
1	0.876145		a_6 * 0.55 + b		2.176086
1	4.729183		a_7 * 0.88 + b		6.966537
1	1.834105		a_8 * 4.73 + b		3.784180
1	2.904804		a_9 * 1.83 + b		2.281207

Mnożenie macierzy (reprezentacja regresji)

$$y = AX + b, \quad b = 0$$

$$y = AX$$

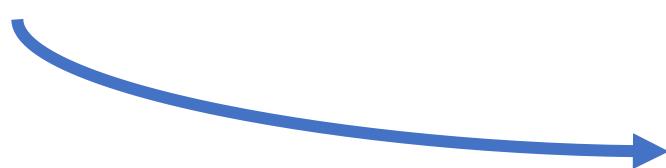
	długość	szerokość	wysokość
	79.564798	139.113256	165.247195
	67.174099	47.367705	286.349772
	118.450682	10.869266	83.118642
	14.156648	25.846663	189.236549
	200.551913	95.144694	37.003291
	63.339889	99.574222	137.854002
	156.675029	204.190017	16.619945

	opóźnienie
	0.601009
	3.031186
	4.160437
	1.326778
	3.453975
	1.106210
	0.091691

	współczynniki
	a_1
	a_2
	a_3

	współczynniki
	a_1
	a_2
	a_3

$$\text{opóźnienie} = a_1 \times \text{długość} + a_2 \times \text{szerokość} + a_3 \times \text{wysokość}$$



$$y = AX =$$

	długość	szerokość	wysokość	opóźnienie
	79.564798	139.113256	165.247195	0.601009
	67.174099	47.367705	286.349772	3.031186
	118.450682	10.869266	83.118642	4.160437
	14.156648	25.846663	189.236549	1.326778
	200.551913	95.144694	37.003291	3.453975
	63.339889	99.574222	137.854002	1.106210
	156.675029	204.190017	16.619945	0.091691

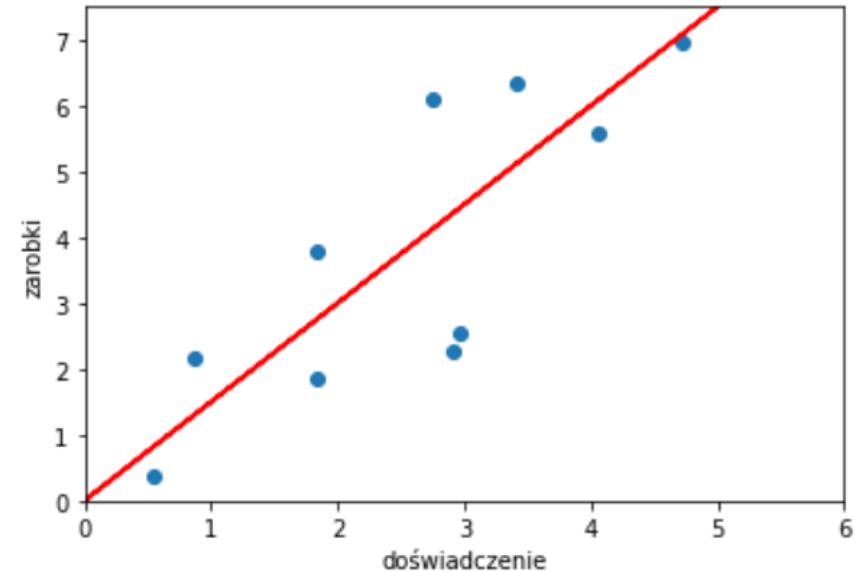
Metoda najmniejszych kwadratów

$$Y = \beta X + e$$

$$\hat{Y} = \beta X$$

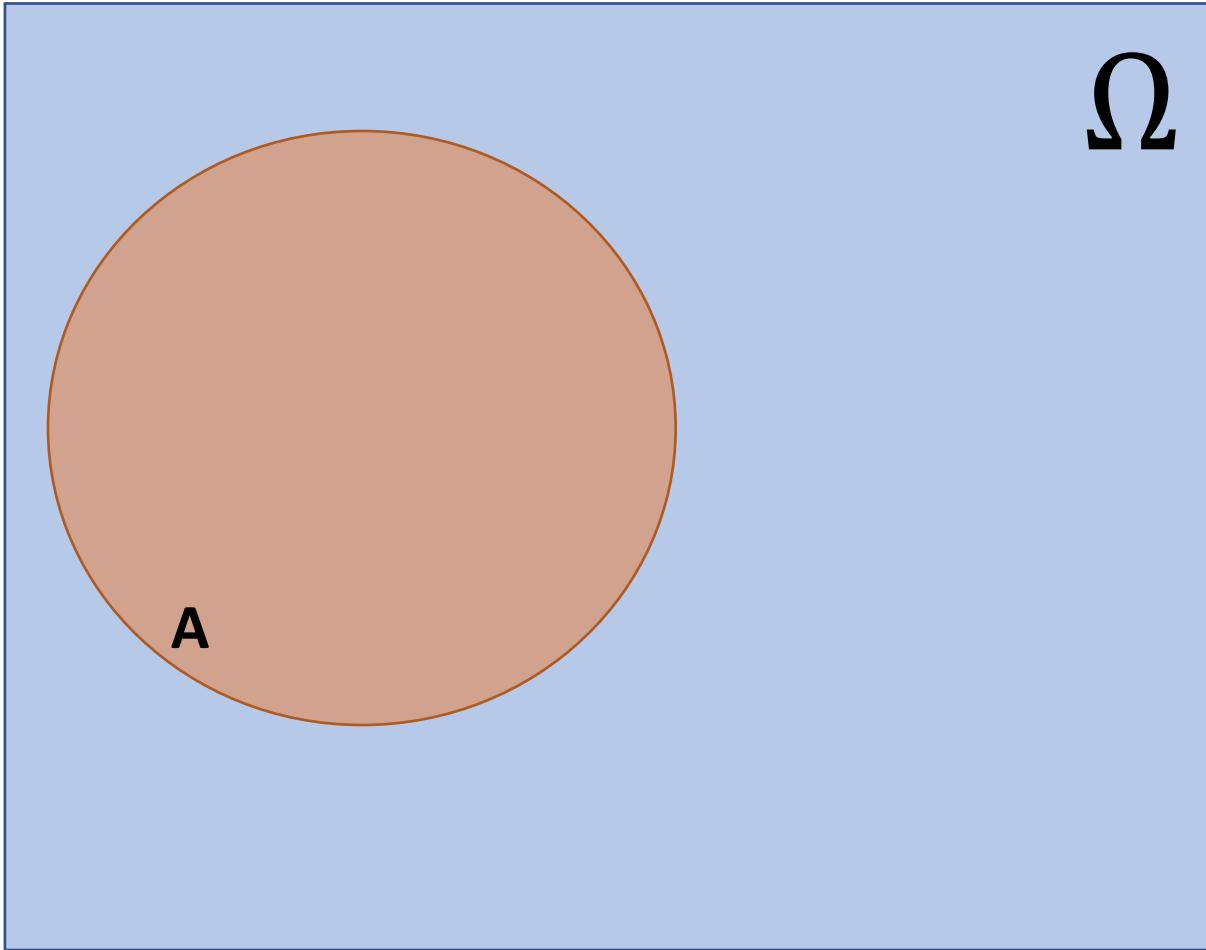
$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

$$\mathbf{X} = \begin{pmatrix} x_{1,0} & x_{1,1} & \dots & x_{1,p-1} \\ x_{2,0} & x_{2,1} & \dots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,0} & x_{n,1} & \dots & x_{n,p-1} \end{pmatrix}.$$



$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

Prawdopodobieństwo



Ω

A

$$P(A) = \frac{|A|}{|\Omega|}$$

$$\mathbb{P}(\emptyset) = 0,$$

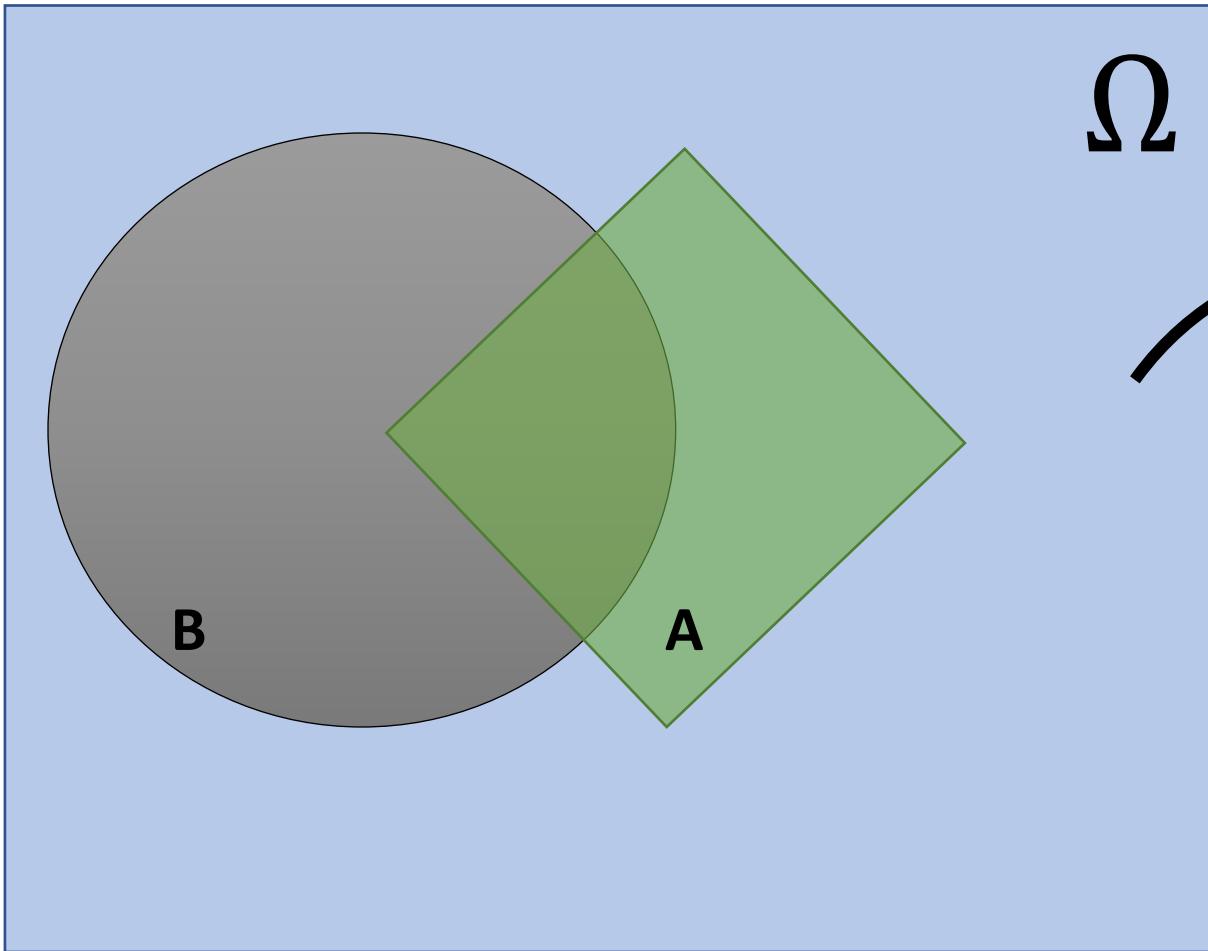
$$A \subset B \rightarrow \mathbb{P}(A) \leq \mathbb{P}(B),$$

$$0 \leq \mathbb{P}(A) \leq 1,$$

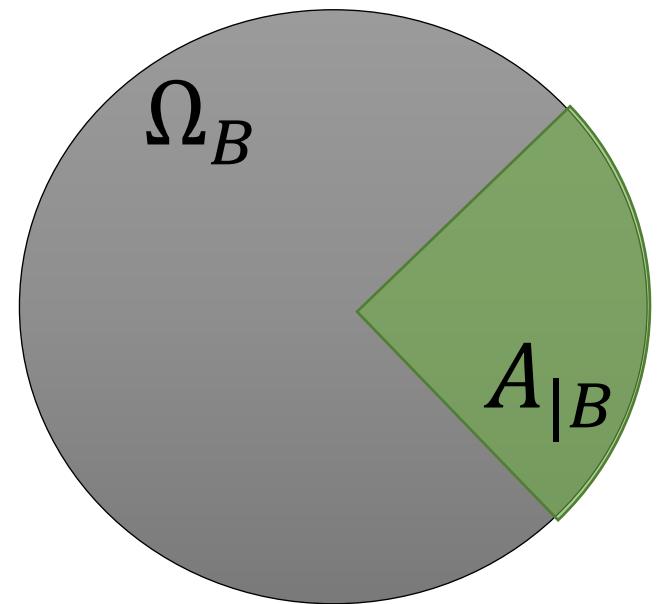
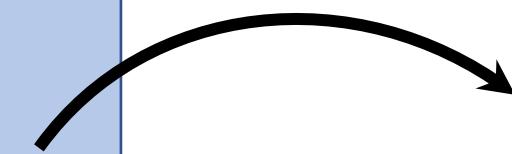
$$\mathbb{P}(A^C) = 1 - \mathbb{P}(A),$$

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

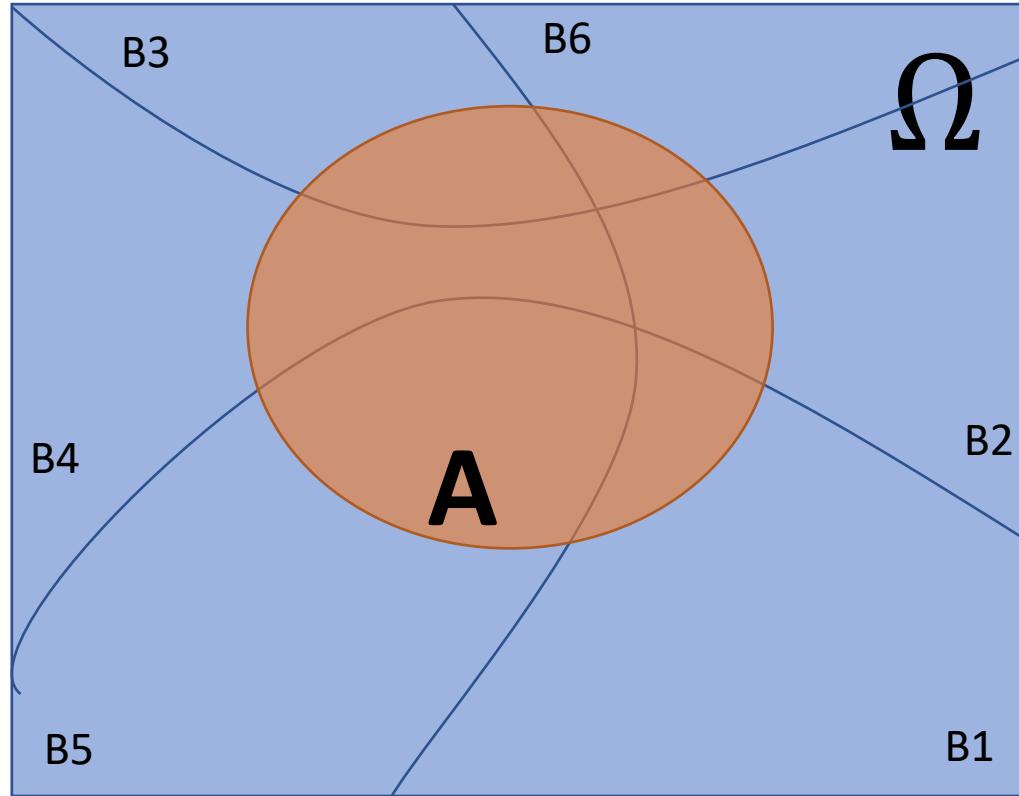
Prawdopodobieństwo warunkowe



$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$



Prawdopodobieństwo całkowite



$$\begin{aligned}P(A) = & P(A|B1)P(B1) \\& + P(A|B2)P(B2) \\& + P(A|B3)P(B3) \\& + P(A|B4)P(B4) \\& + P(A|B5)P(B5) \\& + P(A|B6)P(B6)\end{aligned}$$

$$P(\Omega) = P(B1) + P(B2) + P(B3) + P(B4) + P(B5) + P(B6)$$

Twierdzenie Bayesa

przykład praktyczny (prawdopodobieństwo choroby)

Dany jest test o wrażliwości 80%, co oznacza, że jeśli choroba występuje to test będzie pozytywny z prawdopodobieństwem 0.8.

Prawdopodobieństwo wystąpienia choroby to 0.004.

Test zwraca 10% fałszywie pozytywnych wyników.

$$P(\text{test}_{\text{pozytywny}} | \text{pacjent}_{\text{chorą}}) = 0.8$$

$$P(\text{pacjent}_{\text{chorą}} | \text{test}_{\text{pozytywny}}) = ?$$

$$P(\text{pacjent}_{\text{chorą}}) = 0.004$$

$$P(\text{test}_{\text{pozytywny}} | \text{pacjent}_{\text{zdrowy}}) = 0.1$$

$$P(\text{pacjent}_{\text{chorą}} | \text{test}_{\text{pozytywny}}) = \frac{P(\text{test}_{\text{pozytywny}} | \text{pacjent}_{\text{chorą}})P(\text{pacjent}_{\text{chorą}})}{P(\text{test}_{\text{pozytywny}})}$$

$$= \frac{P(\text{test}_{\text{pozytywny}} | \text{pacjent}_{\text{chorą}})P(\text{pacjent}_{\text{chorą}})}{P(\text{test}_{\text{pozytywny}} | \text{pacjent}_{\text{chorą}})P(\text{pacjent}_{\text{chorą}}) + P(\text{test}_{\text{pozytywny}} | \text{pacjent}_{\text{zdrowy}})P(\text{pacjent}_{\text{zdrowy}})}$$

$$= \frac{0.8 \times 0.004}{0.8 \times 0.004 + 0.1 \times 0.996} = 0.031$$

Twierdzenie Bayesa

■ $P(\text{test}_{\text{pozytywny}} | \text{pacjent}_{\text{chory}}) = 0.8$

■ $P(\text{pacjent}_{\text{chory}}) = 0.004$

■ $P(\text{test}_{\text{pozytywny}} | \text{pacjent}_{\text{zdrowy}}) = 0.1$

● ilość osób: 10 000

■ ilość osób chorych:

$$10\,000 * P(\text{pacjent}_{\text{chory}}) = 40$$

■ ilość osób chorych z pozytywnym wynikiem:

$$40 * P(\text{test}_{\text{pozytywny}} | \text{pacjent}_{\text{chory}}) = 32$$

■ ilość osób zdrowych z pozytywnym wynikiem:

$$9\,960 * P(\text{test}_{\text{pozytywny}} | \text{pacjent}_{\text{zdrowy}}) = 996$$

■ ilość osób z pozytywnym wynikiem:

$$996 + 32 = 1\,028$$

Jakie jest prawdopodobieństwo wylosowanie osoby chorej wśród wszystkich osób z pozytywnym wynikiem testu?

$$P(\text{pacjent}_{\text{chory}} | \text{test}_{\text{pozytywny}}) = \frac{32}{1028} = \mathbf{0.031}$$

Twierdzenie Bayesa

przykład praktyczny (Naïve Bayes)

$$P(dom|x) = \frac{P(x|dom)P(dom)}{P(x)}$$

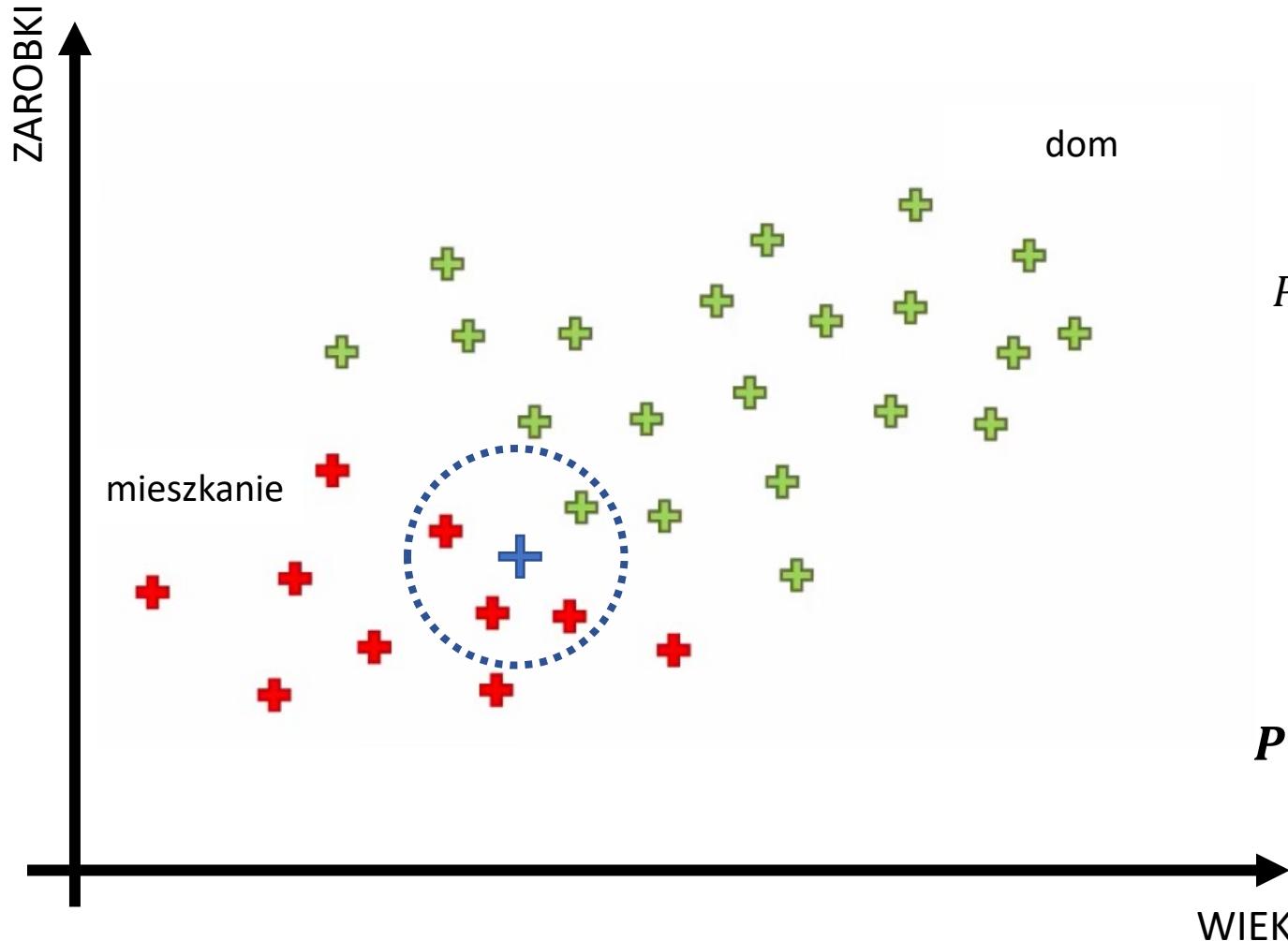
$$P(mieszkanie|x) = \frac{P(x|mieszkanie)P(mieszkanie)}{P(x)}$$

$$P(dom) = \frac{|dom|}{|obserwacje|} = \frac{20}{30}$$

$$P(x|dom) = \frac{|dom w otoczeniu|}{|dom|} = \frac{1}{20}$$

$$\begin{aligned} P(x) &= P(x|dom)P(dom) \\ &\quad + P(x|mieszkanie)P(mieszkanie) \\ &= \frac{1}{20} \frac{20}{30} + \frac{4}{10} \frac{10}{30} = \frac{4}{30} \end{aligned}$$

$$P(dom|x) = 0.25$$



Zmienna Losowa

Zmienna losowa to funkcja:

$$X: \Omega \rightarrow E \subset \mathbb{R}$$

$$\Omega = \{ \text{ ; } \}$$



Funkcja Prawdopodobieństwa (rozkład zmiennej losowej)

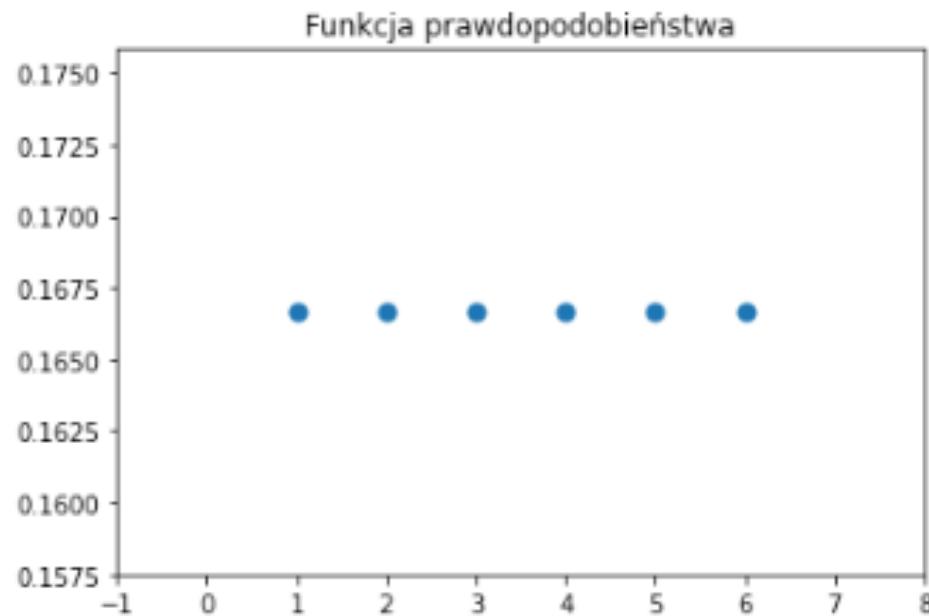
Probability mass function (PMF)

Probability density function (PDF)

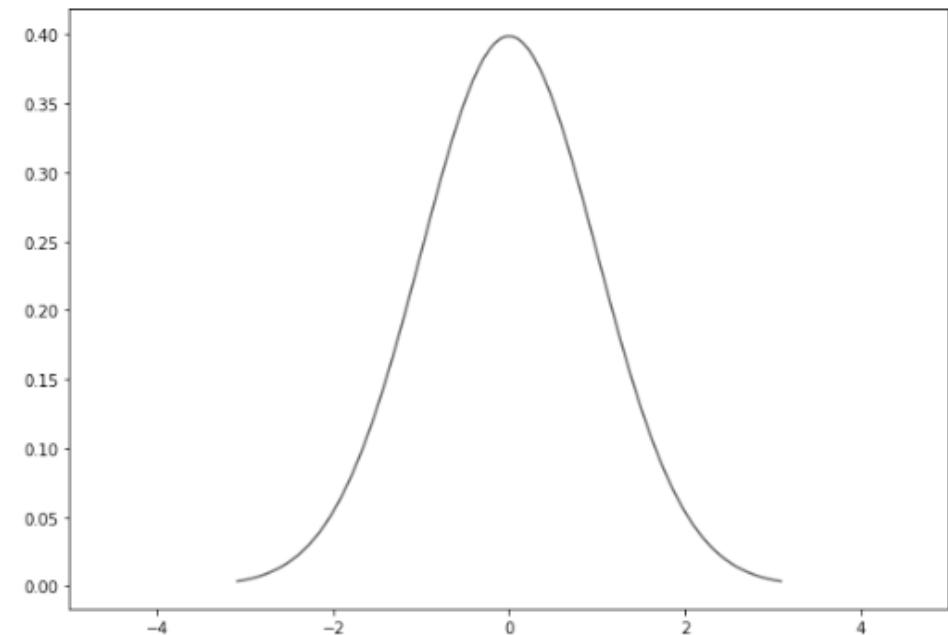
$$f_X(x) = P(X = x)$$

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

$$P(X = 1) = \frac{1}{6} \quad P(X = 2) = \frac{1}{6} \quad P(X = 3) = \frac{1}{6} \quad (\dots)$$



np. rozkład normalny



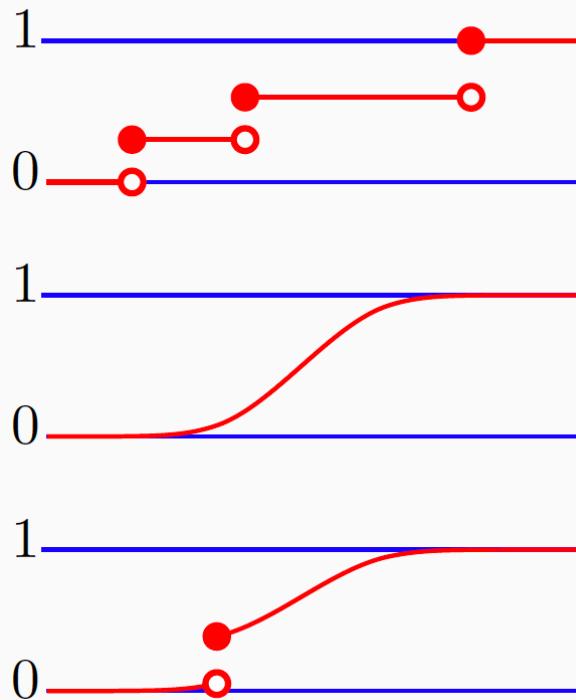
Dystrybuanta

Cumulative distribution function (CDF)

$$F_X(x) = P(X \leq x)$$

Każda dystrybuanta $F(x)$ jest funkcją

- niemalejącą,
- dającą do 1 dla $x \rightarrow +\infty$,
- dającą do 0 dla $x \rightarrow -\infty$,
- prawostronnie ciągłą,
- posiadającą lewostronne granice,
- różniczkowalną prawie wszędzie.



Dystrybuanta

Cumulative distribution function (CDF)

$$F_X(x) = P(X \leq x)$$



$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

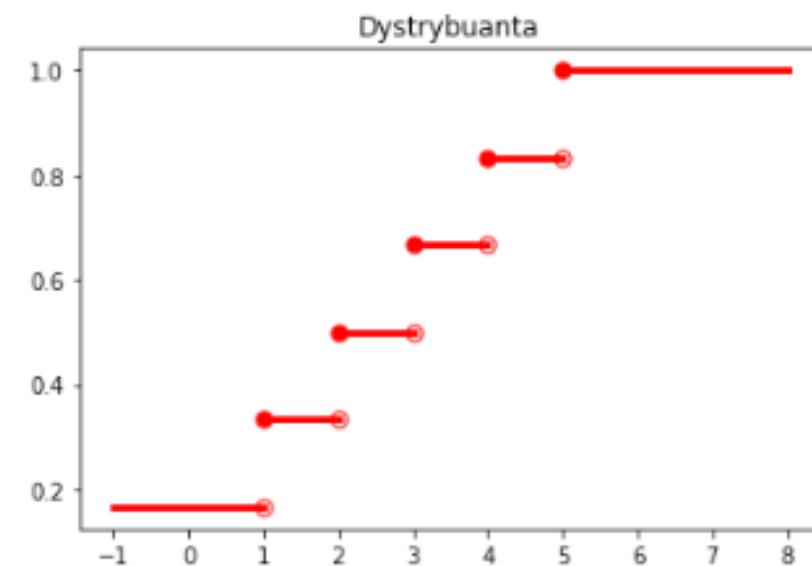
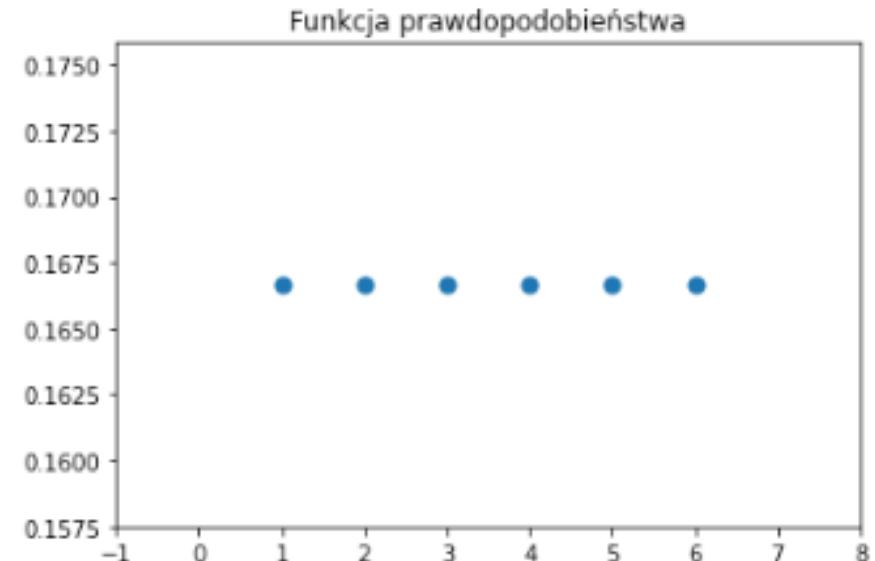
$$P(X = 1) = \frac{1}{6} \quad P(X = 2) = \frac{1}{6} \quad P(X = 3) = \frac{1}{6} \quad (\dots)$$

$$P(X \leq 1) = P(X = 1) = \frac{1}{6}$$

$$P(X \leq 2) = P(X = 1) + P(X = 2) = \frac{2}{6}$$

$$P(X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3) = \frac{3}{6}$$

$$P(X \leq \textcolor{red}{3.42}) = P(X = 1) + P(X = 2) + P(X = 3) = \frac{3}{6} \quad (\dots)$$



Parametry opisujące rozkłady (momenty)

Wartość oczekiwana

Dyskretna zmienna losowa Ciągła zmienna losowa

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i p_i, \quad \mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx.$$

Wartość oczekiwana nazywamy spodziewany wynik doświadczenia losowego przy założonym prawdopodobieństwie jego wystąpienia.

Wariancja

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2\end{aligned}$$

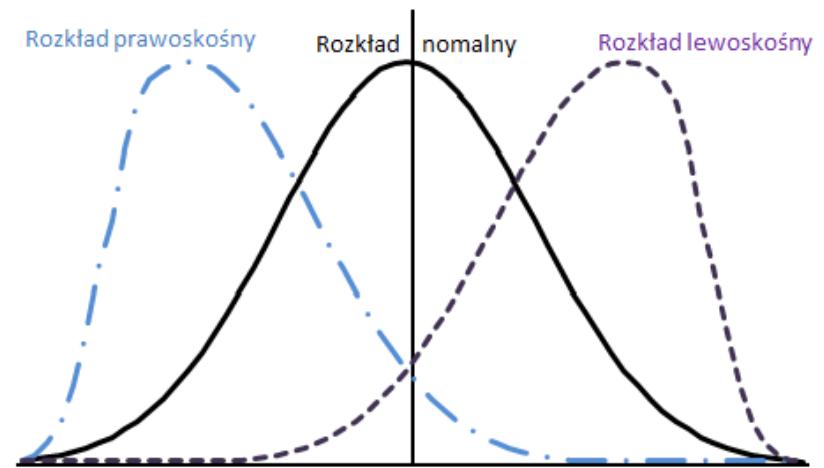
Wariancja jest podstawową miarą zmienności zmiennej losowej. Wariancja informuje o tym, jak duże jest zróżnicowanie wyników w danym zbiorze wyników (zmiennej).

Parametry opisujące rozkłady (momenty)

Skośność

miara symetrii/asymetrii rozkładu. Jeśli rozkład jest idealnie symetryczny, wartość skośności wynosi zero. Z kolei jej wartości ujemne wskazują na rozkład lewoskośny (wydłużone jest lewe ramię rozkładu), a dodatni na prawoskośny (wydłużone jest prawe ramię rozkładu).

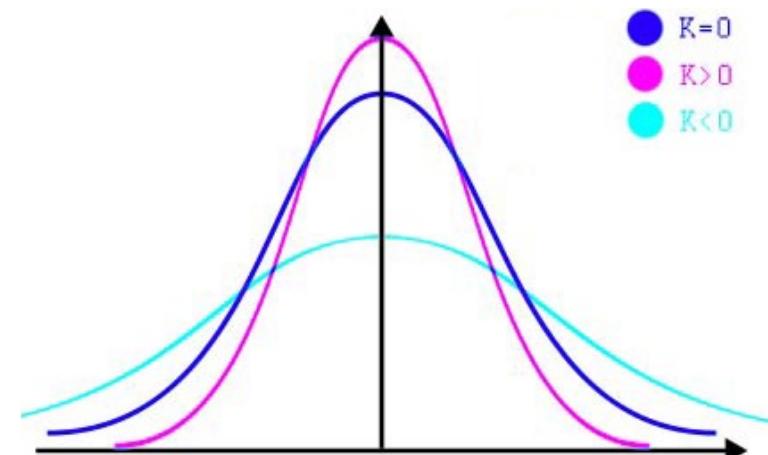
$$\tilde{\mu}_3 = \frac{\sum_i^N (X_i - \bar{X})^3}{(N - 1) * \sigma^3}$$



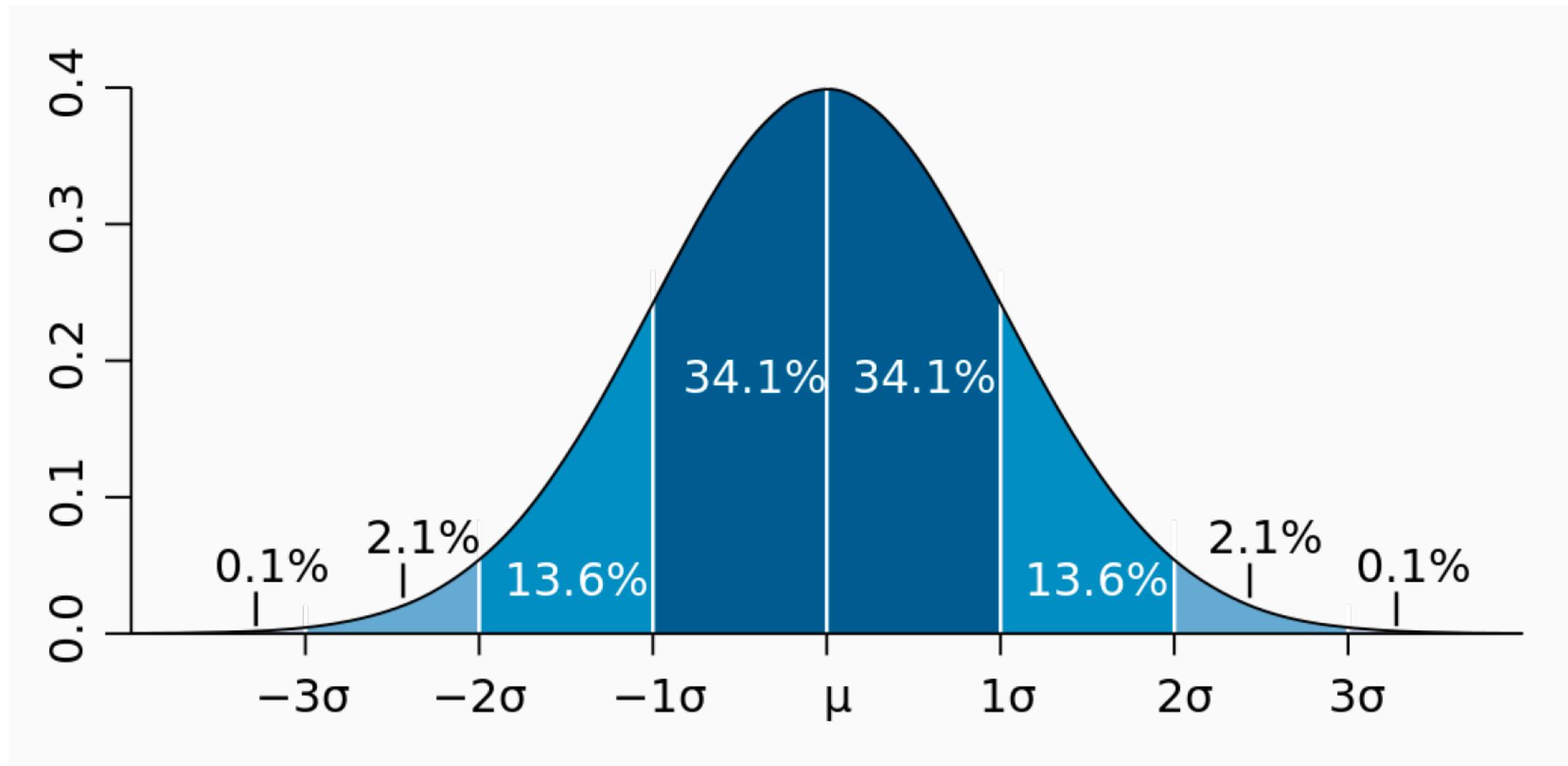
Kurtoza

Względna miara koncentracji i spłaszczenia. Określa rozmieszczenie i koncentrację wartości (zbiorowości) w pobliżu średniej. Występuje on w postaci stosującej moment centralny czwartego rzędu

$$\text{Kurt} = \frac{\mu_4}{\sigma^4}$$

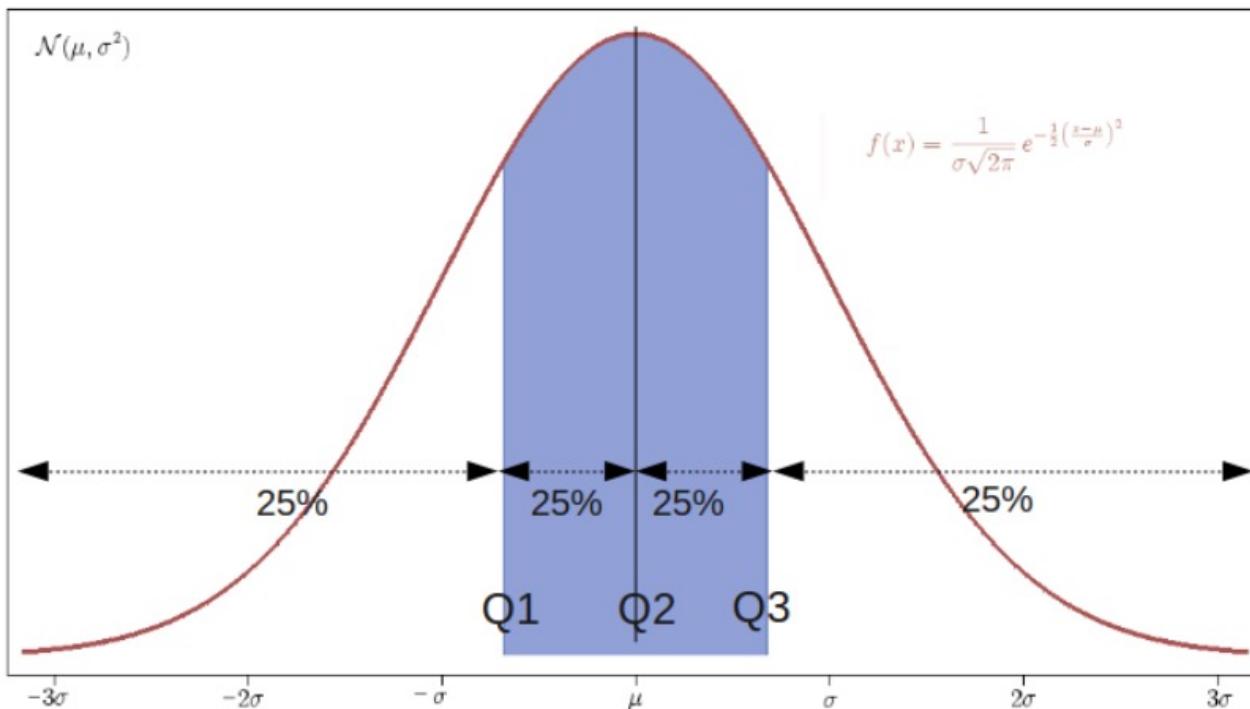


Rozkład normalny – reguła trzech sigm



Kwantyl

Kwantyl rzędu q , ($0 < q < 1$) w populacji jest taką liczbą x_q , że $q \times 100\%$ elementów tej populacji ma wartość badanej cechy nie większą od x_q .



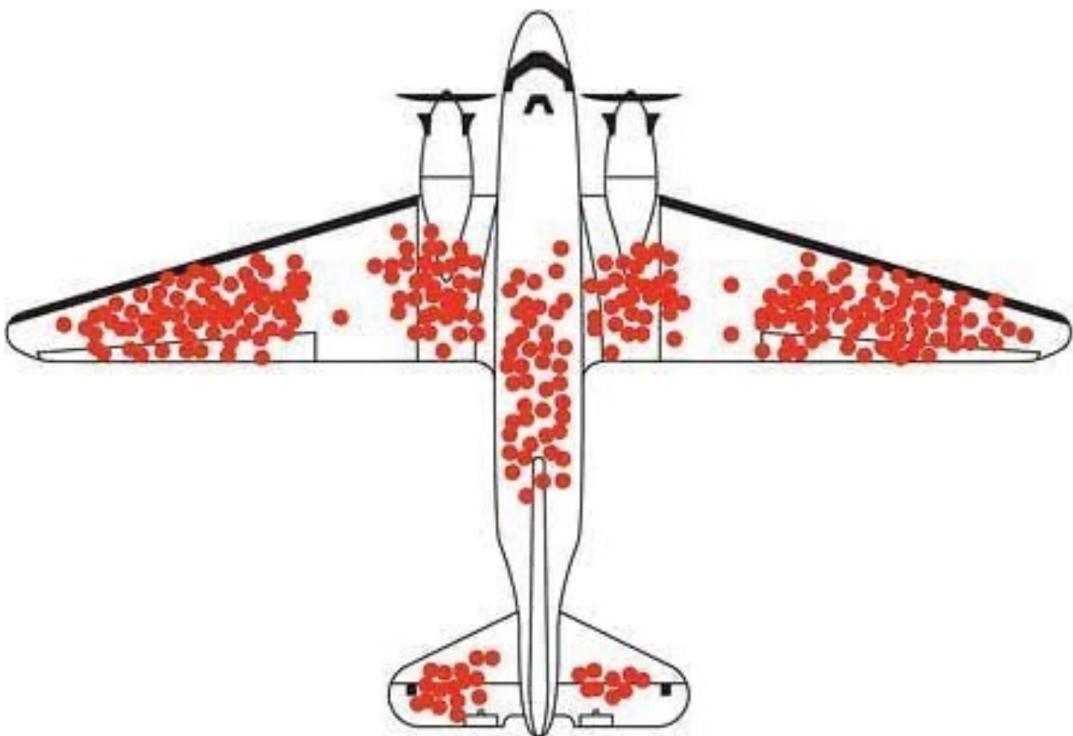
- Kwantyl rzędu $1/2$ to inaczej **mediana**.
- Kwantyle rzędu $1/4, 2/4, 3/4$ są inaczej nazywane **kwartylami**.
- Kwantyle rzędu $1/5, 2/5, 3/5, 4/5$ to inaczej **kwintyle**.
- Kwantyle rzędu $1/10, 2/10, \dots, 9/10$ to inaczej **decyle**.
- Kwantyle rzędu $1/100, 2/100, \dots, 99/100$ to inaczej **percentyle (centyl)**.

Rozstęp kwantylowy - IQR



Case study: *Dopancerznie samolotów*

https://en.wikipedia.org/wiki/Survivorship_bias#In_the_military



The red dots indicate areas of combat damage received by surviving WWII bombers. Where would you add armor to increase survivability? The statistician Abraham Wald recommended reinforcing the areas *without* damage. Since these data came from surviving aircraft only, bombers hit in undotted areas were the ones that did not make it back.

Case study: Produkcja czołgów

An empirical Approach to Economic Intelligence in World War II,
Journal of the American Statistical Association, Vol. 42, No. 237
(Mar., 1947), pp. 72–91.

https://en.wikipedia.org/wiki/German_tank_problem



Produkcia czołgów (dane historyczne):

Miesiąc	est. statystyków	est. agentów	dane prod.
Czerwiec 1940	169	1000	122
Czerwiec 1941	244	1550	271
Sierpień 1942	327	1550	342

Paradoks Simpsona

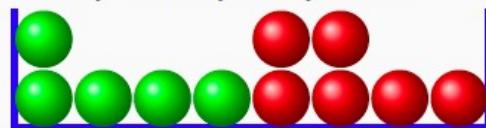
Przyjęci na studia (Univ. of California, Berkeley)

	Zgłoszeń	Przyjętych
Mężczyźni	2691	(1198) 45%
Kobiety	1835	(614) 33%

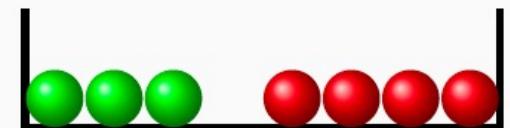
Przyjęci z podziałem na kierunki

K.	Mężczyźni		Kobiety	
	Zgłoszeń	Przyjętych	Zgłoszeń	Przyjętych
A	825	(512) 62%	108	(89) 82%
B	560	(353) 63%	25	(17) 68%
C	325	(120) 37%	593	(219) 37%
D	417	(138) 33%	375	(131) 35%
E	191	(53) 28%	393	(134) 34%
F	373	(22) 6%	341	(24) 7%

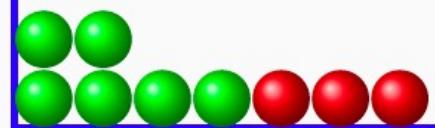
Wybieramy koszyk, z koszyka losujemy kulę, wygrywa zielona.
Który koszyk wybrać?



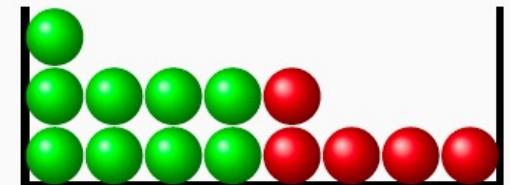
$$\frac{5}{11} > \frac{3}{7}$$



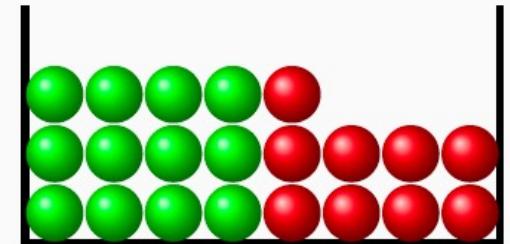
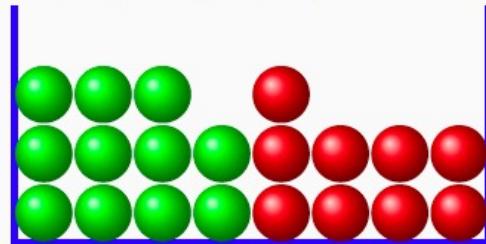
A teraz?



$$\frac{6}{9} > \frac{9}{14}$$



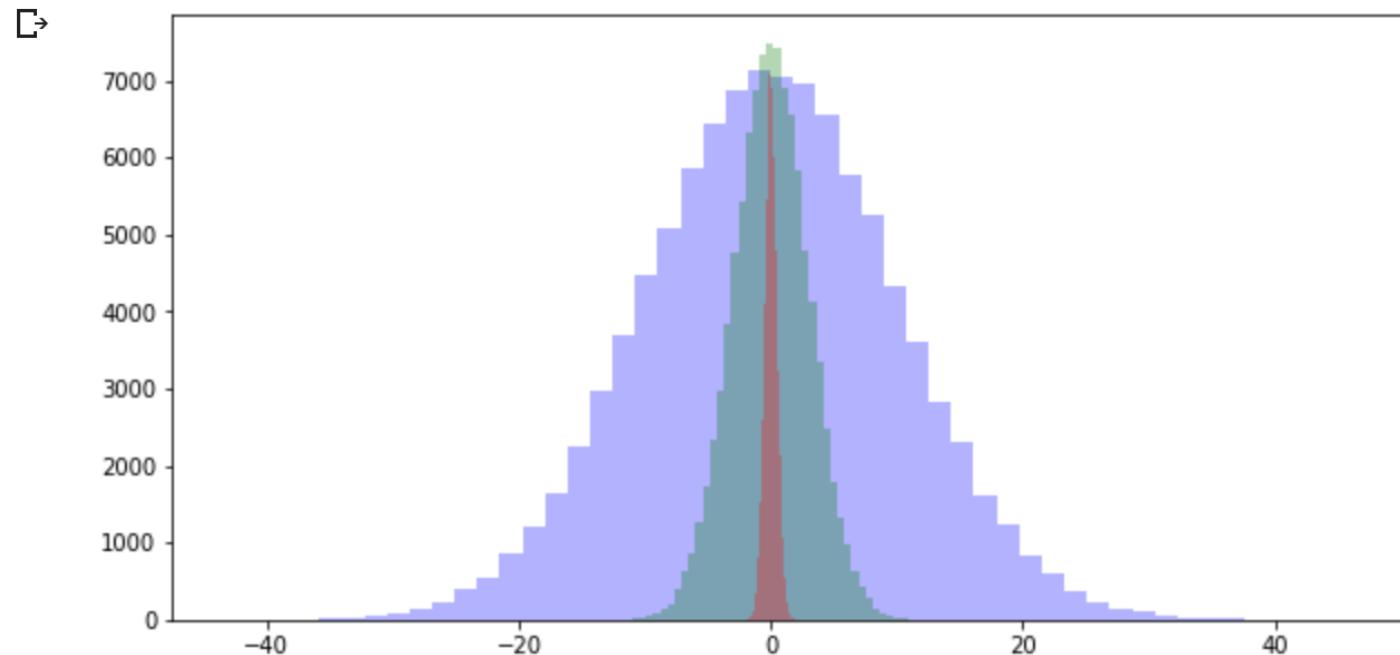
A jak połączymy zawartość koszyków?



Wariancja

✓ 0s

```
N1 = np.random.normal(loc=0.0, scale= 0.5, size=100_000)
N2 = np.random.normal(loc=0.0, scale= 3.0, size=100_000)
N3 = np.random.normal(loc=0.0, scale=10.0, size=100_000)
plt.figure(figsize=(10, 5))
plt.hist(N3, bins=50, alpha=0.3, color='blue')
plt.hist(N2, bins=50, alpha=0.3, color='green')
plt.hist(N1, bins=50, alpha=0.3, color='red')
plt.show()
```



Podatawowe statystyki – wzory!

wartość oczekiwana:

$$\mathbb{E}X = \sum_{i=1}^n x_i p_i$$

wariancja:

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2\end{aligned}$$

odchylenie standardowe:

$$\begin{aligned}\sigma(X) &= \sqrt{\mathbb{E}[(X - \mathbb{E}[X])^2]} \\ &= \sqrt{\mathbb{E}[X^2] - (\mathbb{E}[X])^2}\end{aligned}$$

kowariancja: $\text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X) \cdot (Y - \mathbb{E}Y)]$

Współczynnik

korelacji Pearsona :

$$r_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

statystyka:

Wariancja populacji:

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

Wariancja próbki:

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

Wariancja-znana średnia:

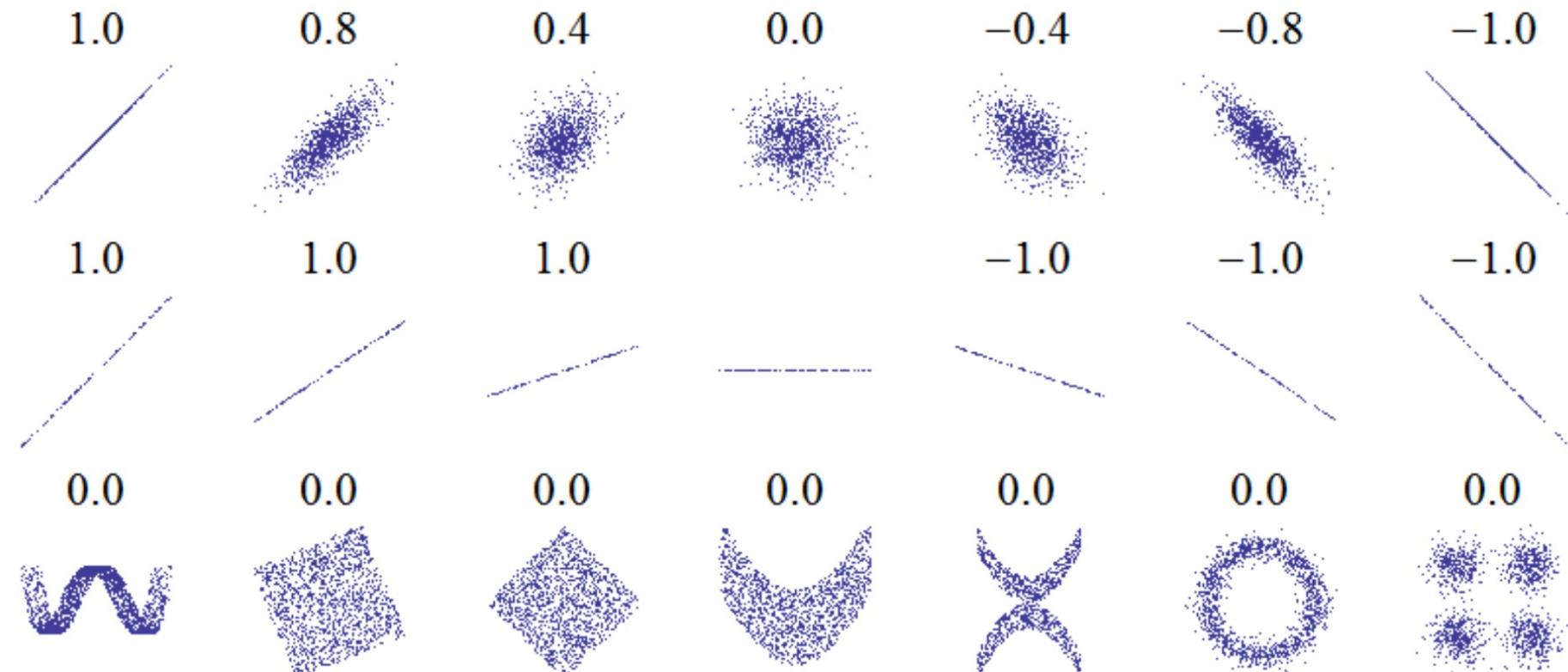
$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

\bar{x} – średnia

μ – wartość oczekiwana

Współczynnik korelacji Pearsona

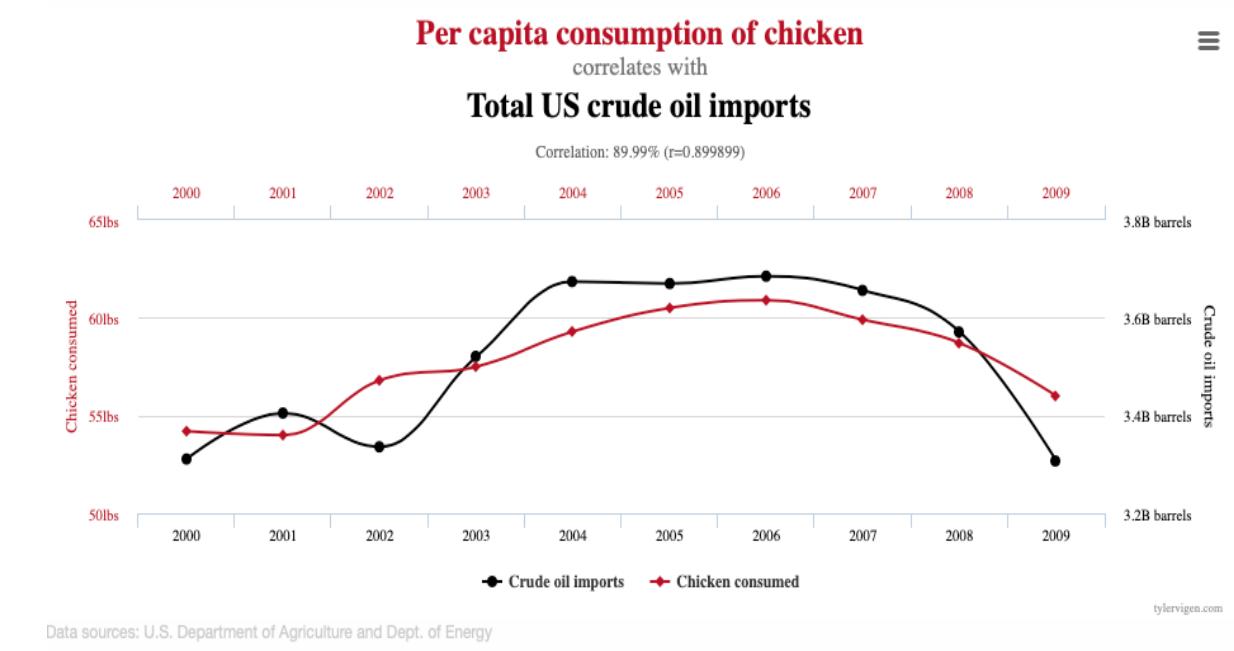
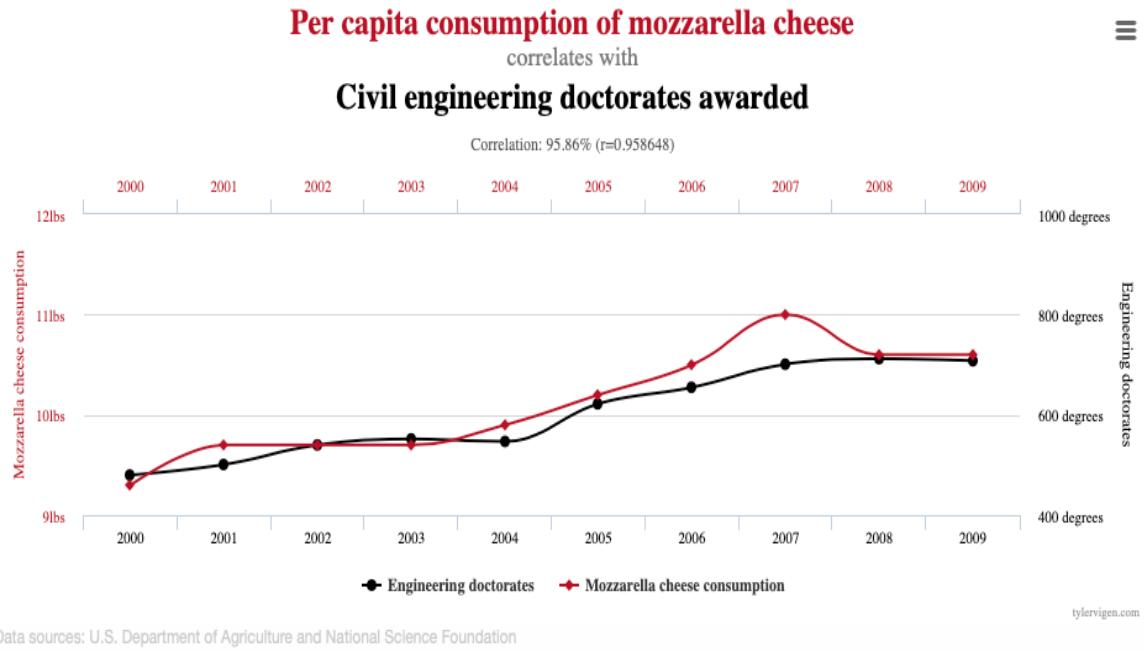
współczynnik określający poziom zależności liniowej między zmiennymi losowymi



Korelacja

to nie przyczynowość!!!
f o u!6 b1sλcsλuomoṣćiii

Korelacja to nie przyczynowość!!!



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

Testy statystyczne

Mamy do czynienia z dwoma hipotezami:

- **H₀** hipotezą zerową, czyli hipotezą, którą w pewnym sensie poddajemy weryfikacji,
- **H₁** hipotezą alternatywną, sprzeczną z hipotezą zerową, ale niekoniecznie ją dopełniającą.

Celem jest wskazanie właściwej hipotezy minimalizując przy tym prawdopodobieństwo popełnienia błędu.

Definicja

Odrzucenie hipotezy H_0 , gdy jest ona prawdziwa nazywamy **błędem I rodzaju**. Prawdopodobieństwo popełnienia takiego błędu na podstawie pewnego testu nazywane jest **poziomem istotności** tego testu i oznaczane najczęściej symbolem α .

Definicja

Przyjęcie hipotezy H_0 , gdy w rzeczywistości nie jest ona prawdziwa nazywamy **błędem II rodzaju**.

Prawdopodobieństwo popełnienia takiego błędu na podstawie pewnego testu oznaczane jest najczęściej symbolem β . Wielkość $1 - \beta$ nazywana jest **mocą testu**.

Testy statystyczne - rodzaje błędów

Obiekt obserwowany w systemie radarowym może podlegać testowi z hipotezami:

H₀ – obiekt jest pociskiem odrzutowym,

H₁ – obiekt jest pasażerskim odrzutowcem.

Błąd I rodzaju polega na zignorowaniu zagrożenia (potraktowanie pocisku jako odrzutowca),

Błąd II rodzaju polega na fałszywym alarmie (mogącym się zakończyć zestrzeleniem samolotu pasażerskiego).

Na podstawie pewnego testu medycznego weryfikuje, czy pacjent cierpi na schorzenie X:

H₀ – nie (negatywny wynik testu),

H₁ – tak, cierpi (pozytywny wynik testu),

Błąd I rodzaju polega na „fałszywym alarmie” (podjęcie niepotrzebnego leczenia mogącego mieć negatywne skutki uboczne),

Błąd II rodzaju polega na zignorowaniu zagrożenia (nie podjęcie terapii gdy jest ona potrzebna).

p-value

Istotność statystyczna wyniku testu, p-wartość, wartość p

Definicja

p-wartością dla zaobserwowanej próby nazywamy minimalną wartość poziomu istotności α , dla której hipoteza zerowa byłaby odrzucona.

Definicja

Odrzucenie hipotezy H_0 , gdy jest ona prawdziwa nazywamy błędem I rodzaju. Prawdopodobieństwo popełnienia takiego błędu na podstawie pewnego testu nazywane jest poziomem istotności tego testu i oznaczane najczęściej symbolem α .

- Badamy pewne zjawisko, które wiemy na 100% że nie zachodzi.
- Wybieramy dwie grupy z populacji.
- Jedną z tych grup poddajemy zjawisku
- Dostajemy rezultat:

średnia = 38; odchylenie standardowe = 4

średnia = 32; odchylenie standardowe = 3,8

H_0 – nie ma wpływu zjawiska

H_1 – zjawisko zachodzi

$p = 0,840$

- Prawdopodobieństwo otrzymania tak dużej różnicy wynosi jeśli hipoteza zerowa faktycznie jest prawdziwa, wynosi aż 0.84.

Test normalności Shapiro-Wilka

Test Shapiro-Wilko jest uznawany za najlepszy test do sprawdzenia normalności rozkładu zmiennej losowej. Głównym atutem tego testu jest jego duża moc, tzn. dla ustalonego α prawdopodobieństwo odrzucenia hipotezy H_0 , jeśli jest ona fałszywa, jest większe niż w przypadku innych tego typu testów.

H_0 : *Rozkład badanej cechy jest rozkładem normalnym.*

H_1 : *Rozkład badanej cechy nie jest rozkładem normalnym.*

Testy parametryczne

Testy wymagające założenia normalności rozkładu. Założenie to można sprawdzić np. wykorzystując test normalności Shapiro Wilka. Czasami użycie testów parametrycznych wymaga również innych założeń, jak np. założenie o stałości wariancji w podgrupach w przypadku analizy wariancji (ANOVA).

Test t-studenta dla 1 średniej

Test ten służy weryfikacji hipotezy o równości wartości przeciętnej μ konkretnej liczbie μ_0 .

$$H_0: \mu = \mu_0$$

$$H_1: \begin{array}{l} 1. \mu \neq \mu_0 \\ 2. \mu < \mu_0 \\ 3. \mu > \mu_0 \end{array}$$

Test t-studenta dla 2 średnich

Za pomocą tych testów weryfikujemy hipotezę o równości wartości przeciętnych w dwóch populacjach.

$$H_0: \mu_1 = \mu_2$$

$$H_1: \begin{array}{l} 1. \mu_1 \neq \mu_2 \\ 2. \mu_1 < \mu_2 \\ 3. \mu_1 > \mu_2 \end{array}$$

1. Próby zależne

2. Próby niezależne