Wektor

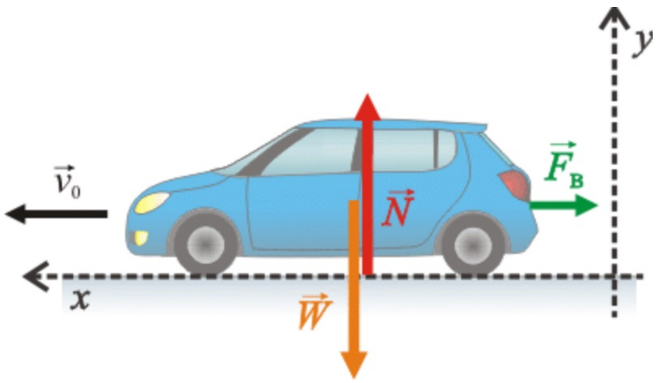
Programista

Tablica wypełniona liczbami

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

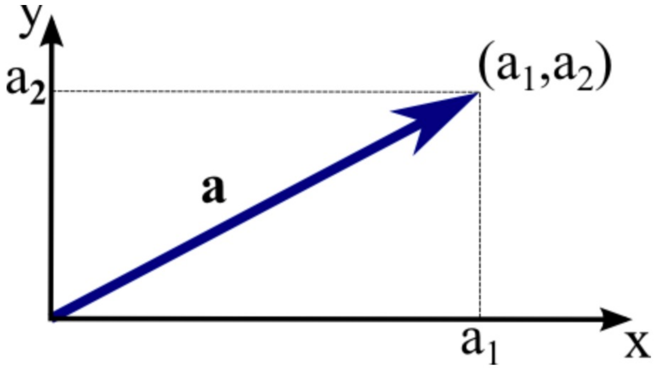
Fizyk

Obiekt opisywany za pomocą długości, kierunku i zwrotu.



Matematyk

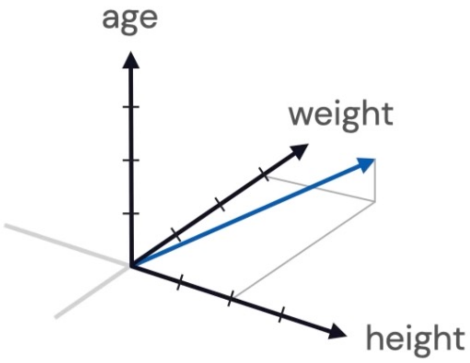
Współrzędne punktu w przestrzeni.



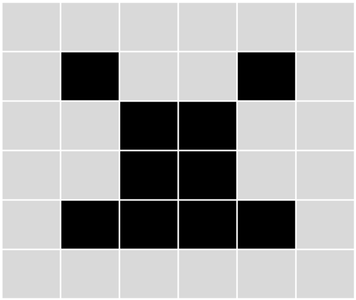
Wektor – reprezentacja danych

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

"p" for "patient"

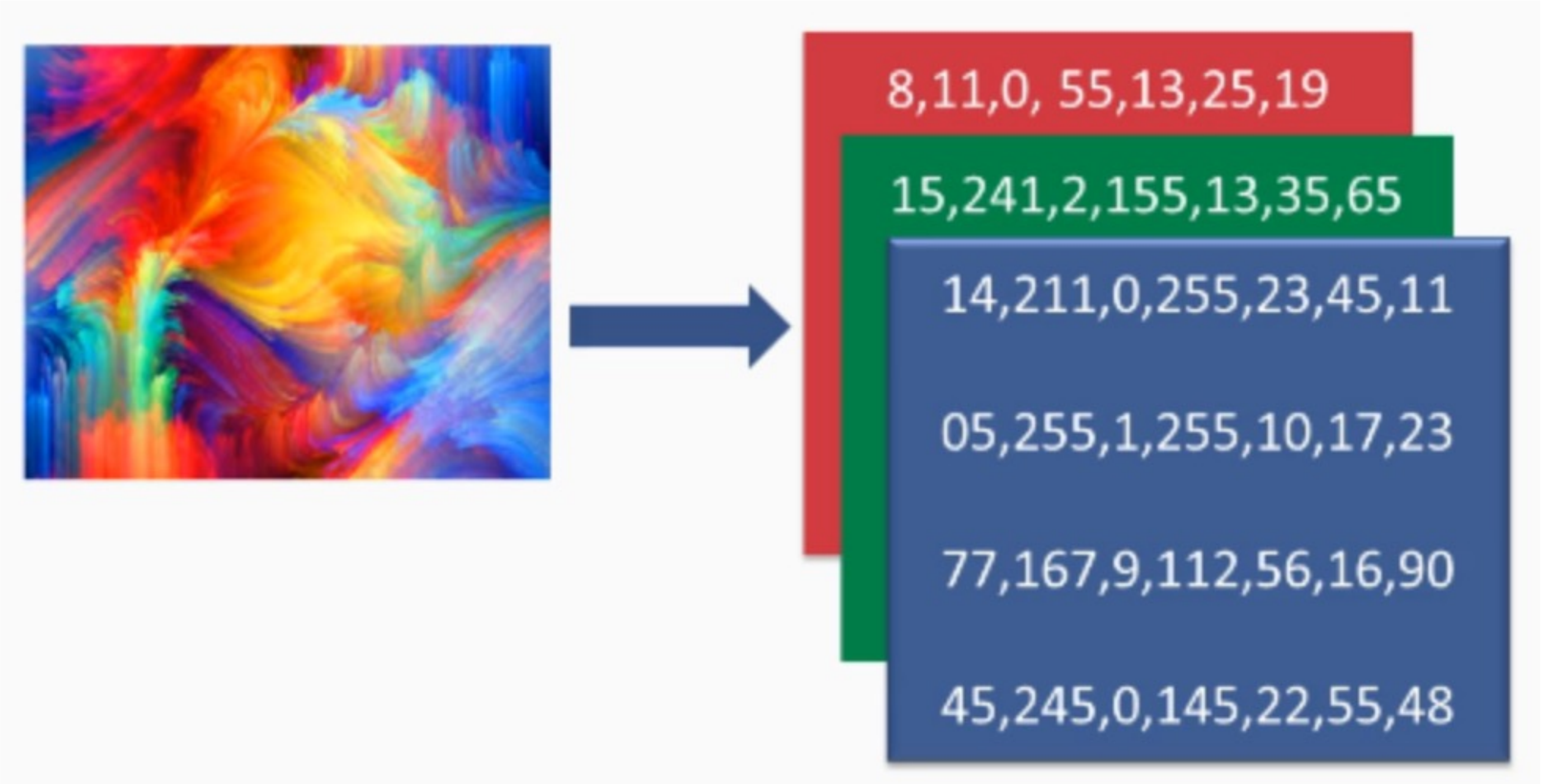


$$cat = \begin{bmatrix} 0 \\ 0 \\ 2 \\ 0 \\ 15 \\ 1 \\ 0 \\ 0 \\ 51 \\ \vdots \\ 0 \end{bmatrix} \begin{array}{l} \text{doc \#1} \\ \text{doc \#2} \\ \text{doc \#3} \\ \text{doc \#4} \\ \text{doc \#5} \\ \text{doc \#6} \\ \text{doc \#7} \\ \text{doc \#8} \\ \text{doc \#9} \\ \\ \text{doc \#1500} \end{array}$$

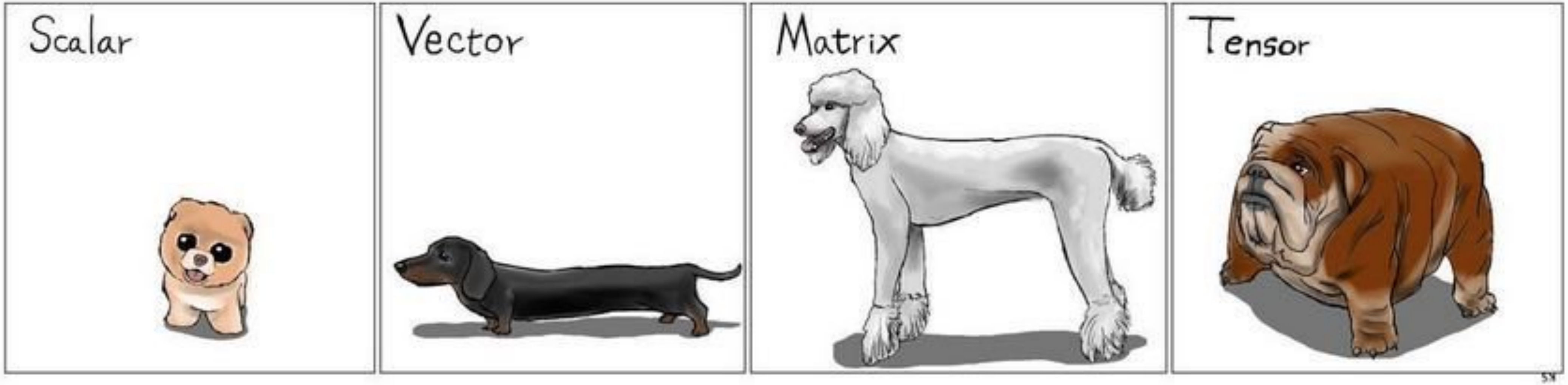


$$\Longrightarrow \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \Longrightarrow \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Tensor - przykład



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 =
```
[[5 6 1 2]
 [8 7 6 3]
 [5 0 6 4]]
```

Z = X * Y =
```
[[15  0  4 18]
 [32 42 30 24]
 [35  0  6 20]]
```

Y =
```
[[3 0 4 9]
 [4 6 5 8]
 [7 0 1 5]]
```

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

$x_{11} = 5 \quad y_{11} = 3 \qquad z_{11} = x_{11} \times y_{11} = 15$

$x_{12} = 6 \quad y_{12} = 0 \qquad z_{12} = x_{12} \times y_{12} = 0$

$X_{3 \times 4}$

$Y_{3 \times 4}$ ↤ Wymiar obu macierzy musi być taki sam

Mnożenie macierzy (algebraiczne) *(dot product)*

X =
```
[[9 2 2]
 [4 0 0]
 [9 3 9]]
```

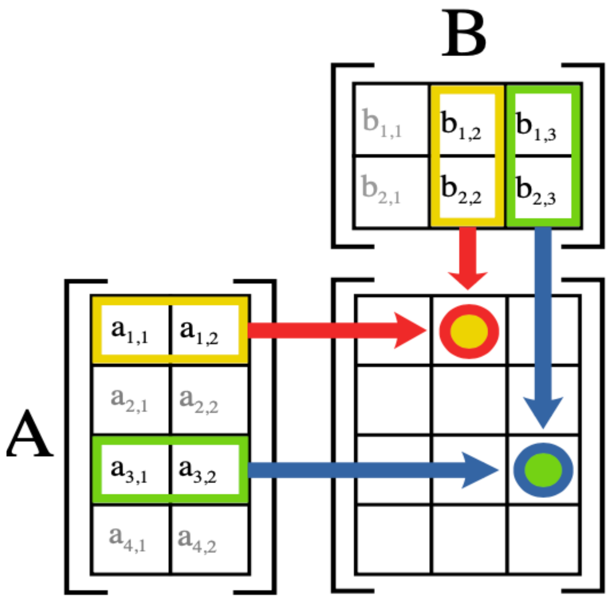
Y =
```
[[8 1 1]
 [9 6 8]
 [7 4 8]]
```

Z = X . Y =
```
[[104  29  41]
 [ 32   4   4]
 [162  63 105]]
```

$z_{11} = x_{11} \times y_{11} + x_{12} \times y_{21} + x_{13} \times y_{31}$ $\quad z_{12} = x_{11} \times y_{12} + x_{12} \times y_{22} + x_{13} \times y_{32}$

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

$A_{p \times m}$



$B_{m \times q}$

↤ *pierwszy wymiar* macierzy **A** musi być taki sam jak *drugi wymiar* macierzy **B**

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

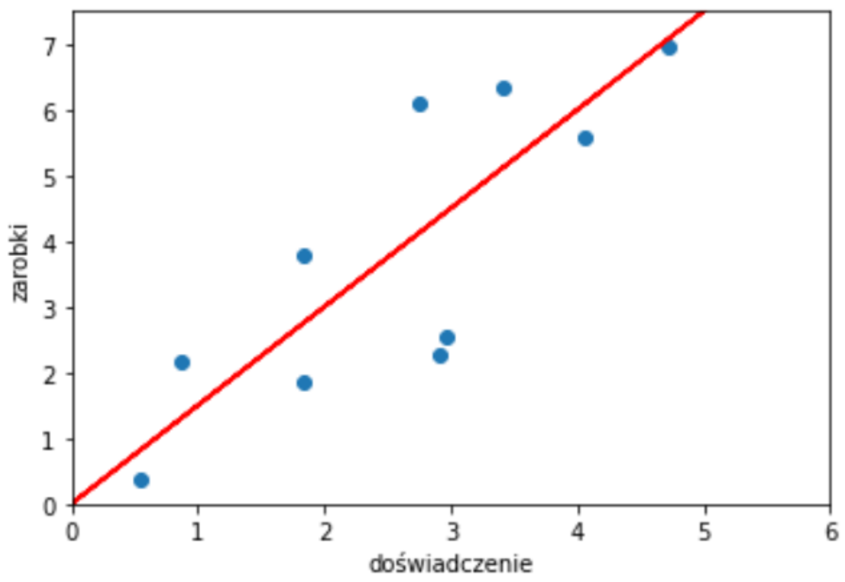
-Morpheus

Mnożenie macierzy (*geometria* regresji)

$$y = AX + b + e$$

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

$$zarobki = \ a \times doświdczenie + b$$



b	doświadczenie	zarobki
1	4.057534	5.586908
1	2.750049	6.105262
1	2.964347	2.533063
1	3.410334	6.332447
1	1.828527	1.840721
1	0.548475	0.355877
1	0.876145	2.176086
1	4.729183	6.966537
1	1.834105	3.784180
1	2.904804	2.281207

$$zarobki_1 = \ a \times doświdczenie_1 + b$$
$$zarobki_2 = \ a \times doświdczenie_2 + b$$
$$zarobki_3 = \ a \times doświdczenie_3 + b$$

(. . .)

b	doświadczenie
1	4.057534
1	2.750049
1	2.964347
1	3.410334
1	1.828527
1	0.548475
1	0.876145
1	4.729183
1	1.834105
1	2.904804

*

współczynniki
b
a

=

a_1 * 4.06 + b
a_2 * 2.75 + b
a_3 * 2.96 + b
a_4 * 3.41 + b
a_5 * 1.83 + b
a_6 * 0.55 + b
a_7 * 0.88 + b
a_8 * 4.73 + b
a_9 * 1.83 + b
a_10 * 2.9 + b

=

zarobki
5.586908
6.105262
2.533063
6.332447
1.840721
0.355877
2.176086
6.966537
3.784180
2.281207

Mnożenie macierzy (reprezentacja regresji)

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

$X =$

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

$y =$

opóźnienie
0.601009
3.031186
4.160437
1.326778
3.453975
1.106210
0.091691

$A =$

współczynniki
a_1
a_2
a_3

$$opóźnienie = a_1 \times długość + \; a_2 \times szerokość + a_3 \times wysokość$$

$y = AX =$

współczynniki
a_1
a_2
a_3

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

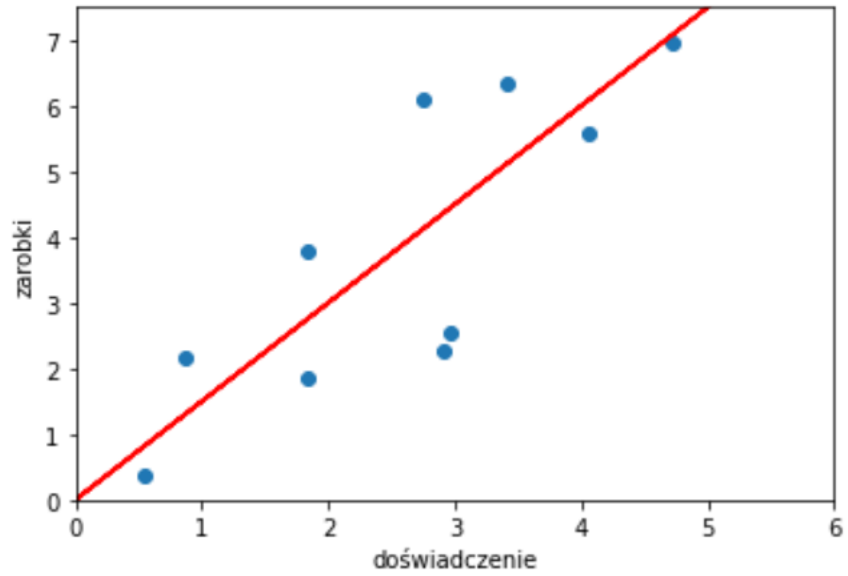
Metoda najmniejszych kwadratów

$$Y = \beta X + e$$

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

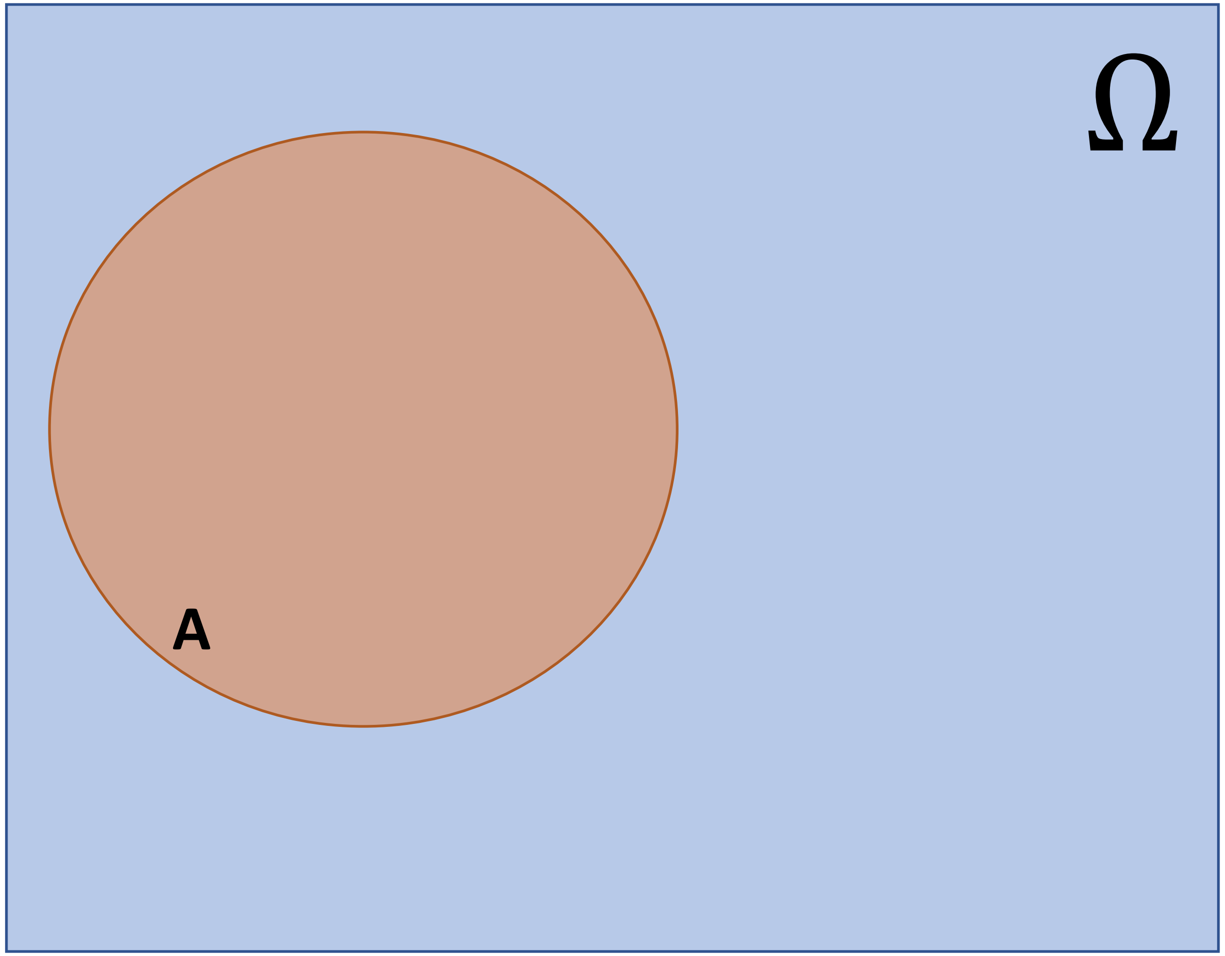
$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

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



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

Prawdopodobieństwo



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

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

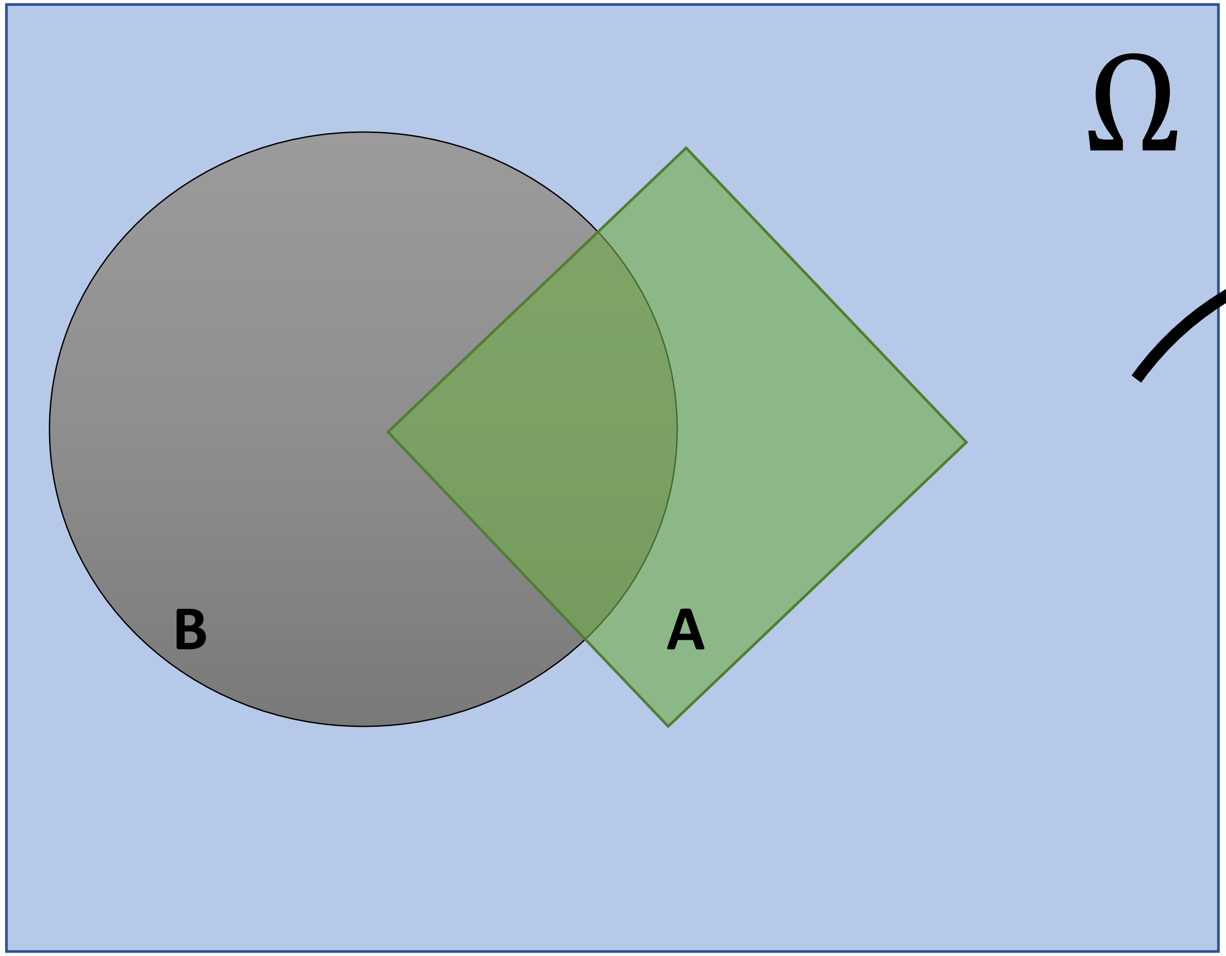
$$A \subset B \rightarrow \mathbb{P}(A) \leq \mathbb{P}(B),$$

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

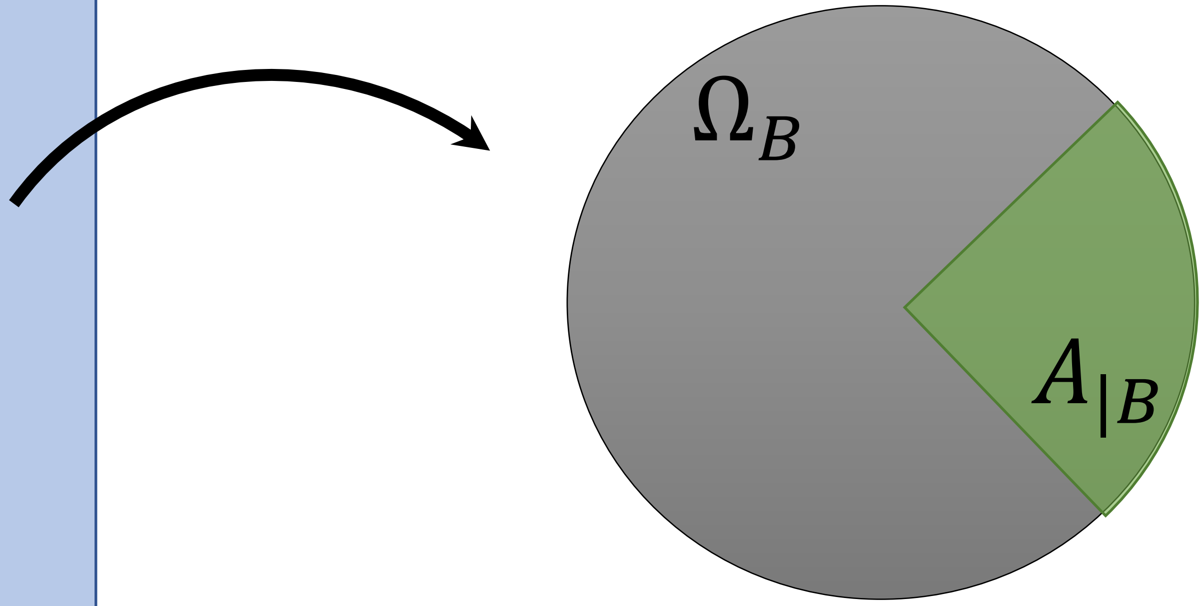
$$\mathbb{P}(A^C) = 1 - \mathbb{P}(A),$$

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

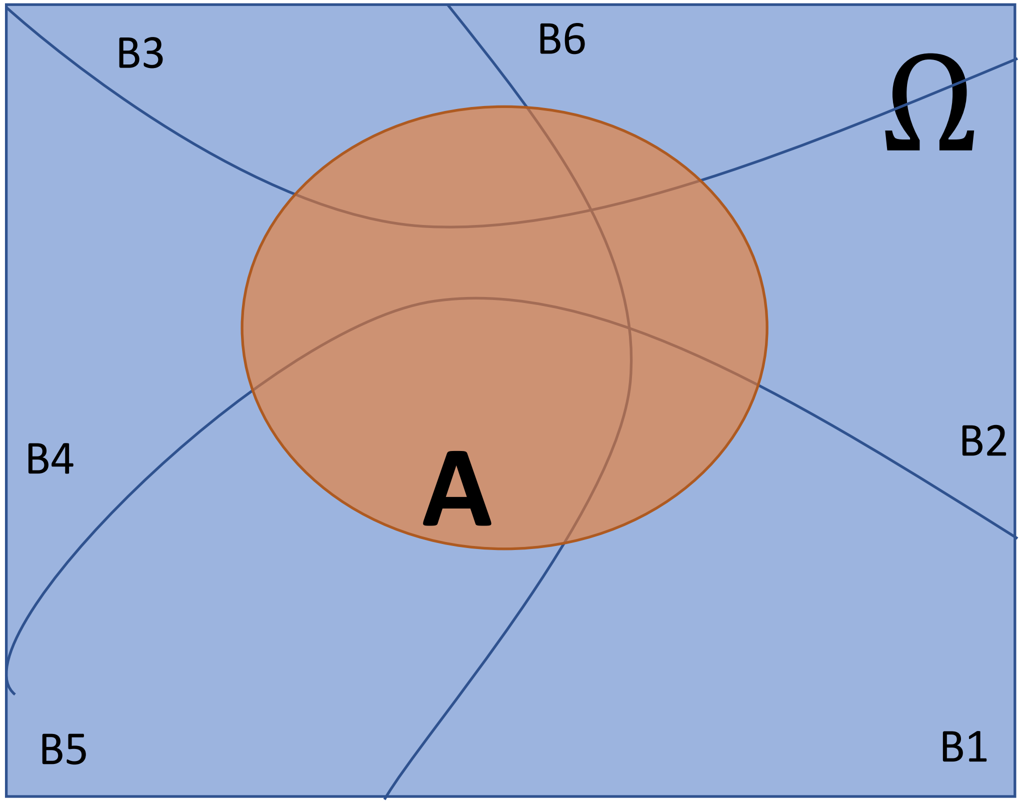
Prawdopodobieństwo warunkowe



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



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.

$$P\left(test_{pozytywny} \middle| pacjent_{chory}\right) = 0.8$$

$$P\left(pacjent_{chory}\right) = 0.004$$

$$P\left(test_{pozytywny} \middle| pacjent_{zdrowy}\right) = 0.1$$

$$P\left(pacjent_{chory} \middle| test_{pozytywny}\right) = ?$$

$$P\left(pacjent_{chory} \middle| test_{pozytywny}\right) = \frac{P\left(test_{pozytywny} \middle| pacjent_{chory}\right) P\left(pacjent_{chory}\right)}{P(test_{pozytywny})}$$

$$= \frac{P\left(test_{pozytywny} \middle| pacjent_{chory}\right) P\left(pacjent_{chory}\right)}{P\left(test_{pozytywny} \middle| pacjent_{chory}\right) P\left(pacjent_{chory}\right) + P\left(test_{pozytywny} \middle| pacjent_{zdrowy}\right) P\left(pacjent_{zdrowy}\right)}$$

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

Twierdzenie Bayesa

$$P(test_{pozytywny}|pacjent_{chory}) = 0.8$$

$$P(pacjent_{chory}) = 0.004$$

$$P(test_{pozytywny}|pacjent_{zdrowy}) = 0.1$$

$ilość\ osób$: $10\ 000$

$ilość\ osób\ chorych$:
$10\ 000\ * P(pacjent_{chory}) = 40$

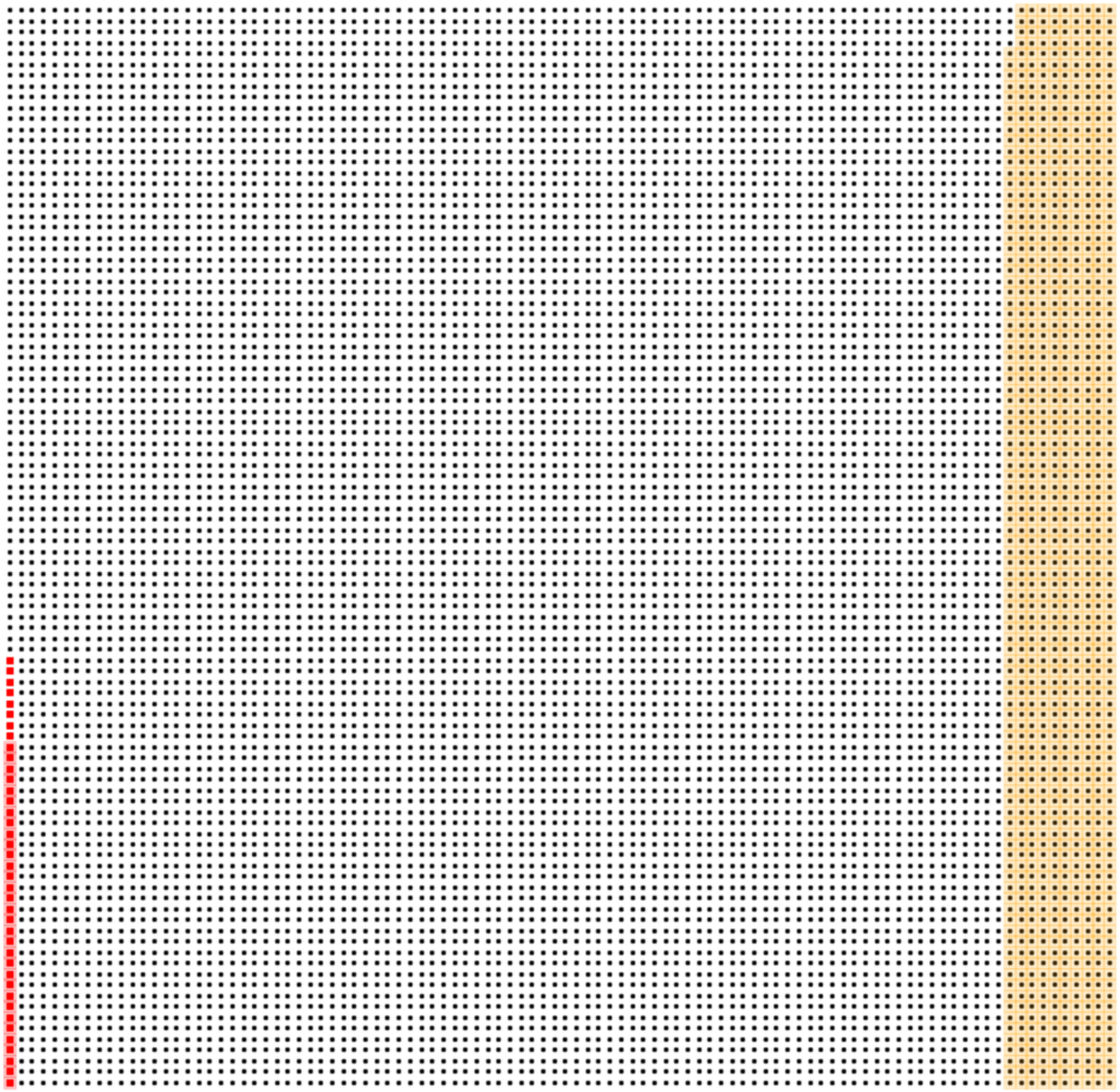
$ilość\ osób\ chorych\ z\ pozytywnym\ wynikiem$:
$40\ *\ P(test_{pozytywny}|pacjent_{chory}) = 32$

$ilość\ osób\ zdrowych\ z\ pozytywnym\ wynikiem$:
$9\ 960\ * P(test_{pozytywny}|pacjent_{zdrowy}) = 996$

$ilość\ osób\ z\ pozytywnym\ wynikiem$:
$996 + 32 = 1\ 028$

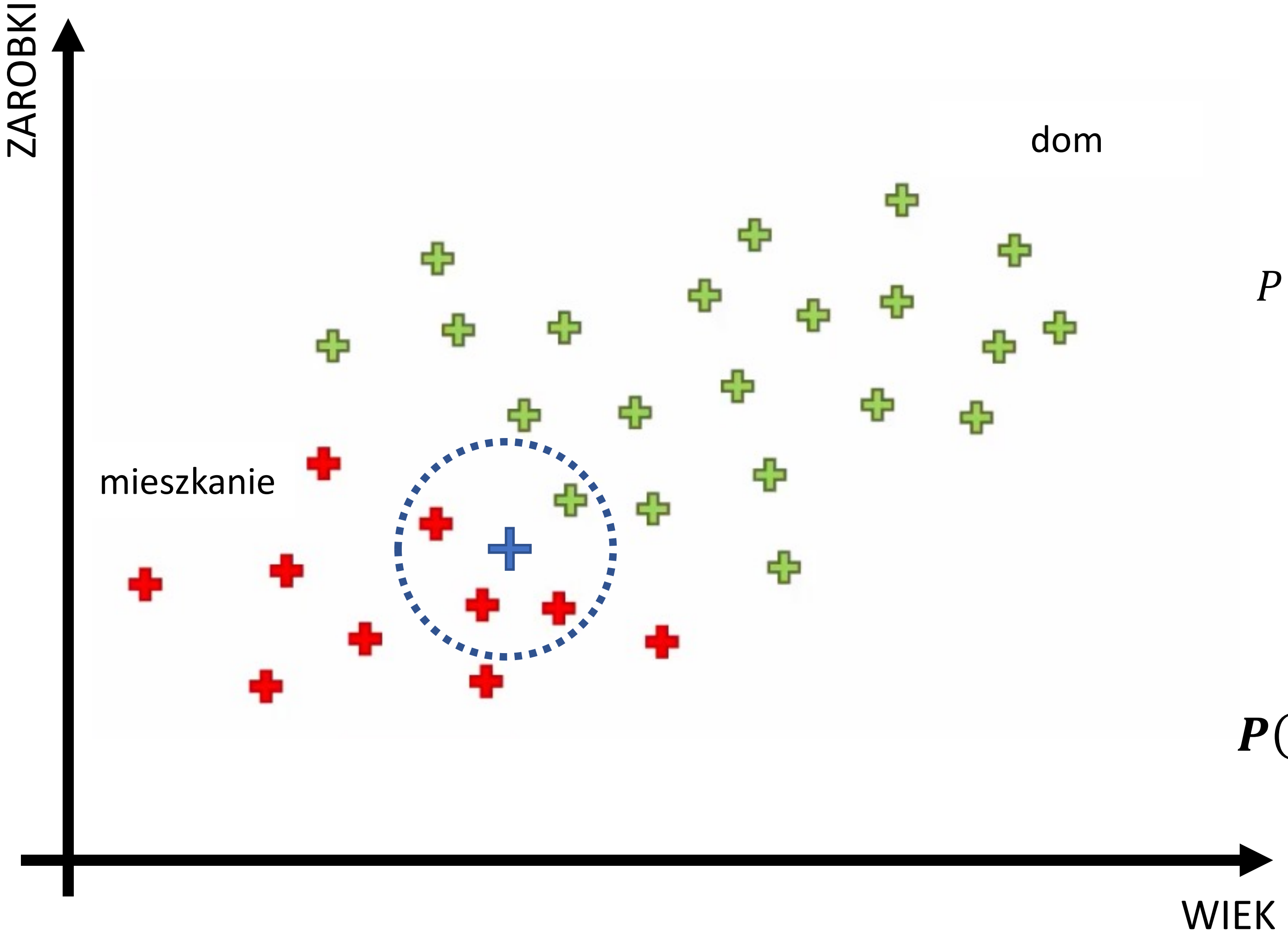
Jakie jest prawdopodobieństwo wylosowanie osoby chorej wśród wszystkich osób z pozytywnym wynikiem testu?

$$\boldsymbol{P(pacjent_{chory}|test_{pozytywny}) = \frac{32}{1028} = 0.031}$$



Twierdzenie Bayesa

przykład praktyczny (Naïve Bayes)



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

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

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

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

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

$$\boldsymbol{P(dom|x) = 0.25}$$

Zmienna Losowa

Zmienna losowa to funkcja:

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

$\Omega = \{$

;

$\}$

$\longrightarrow 1$

$\longrightarrow 0$

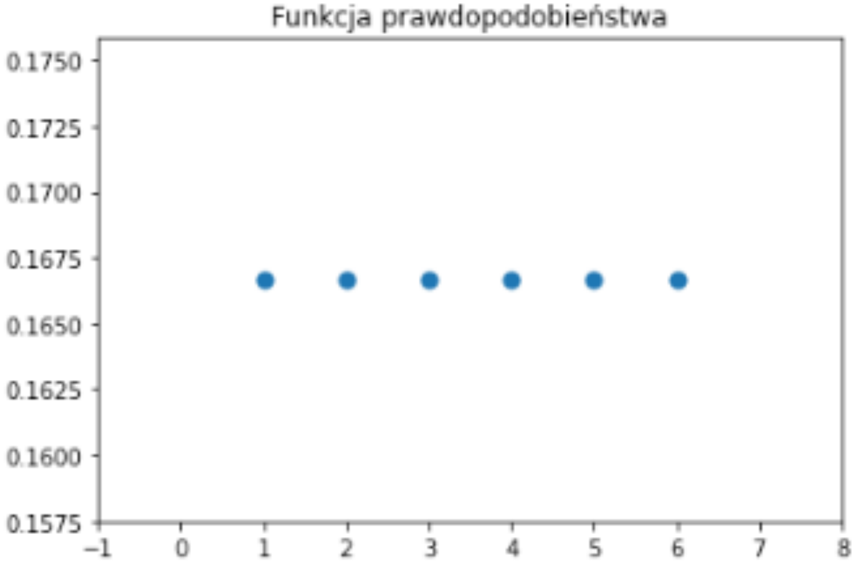
Funkcja Prawdopodobieństwa (rozkład zmiennej losowej)

Probability density function (PDF)

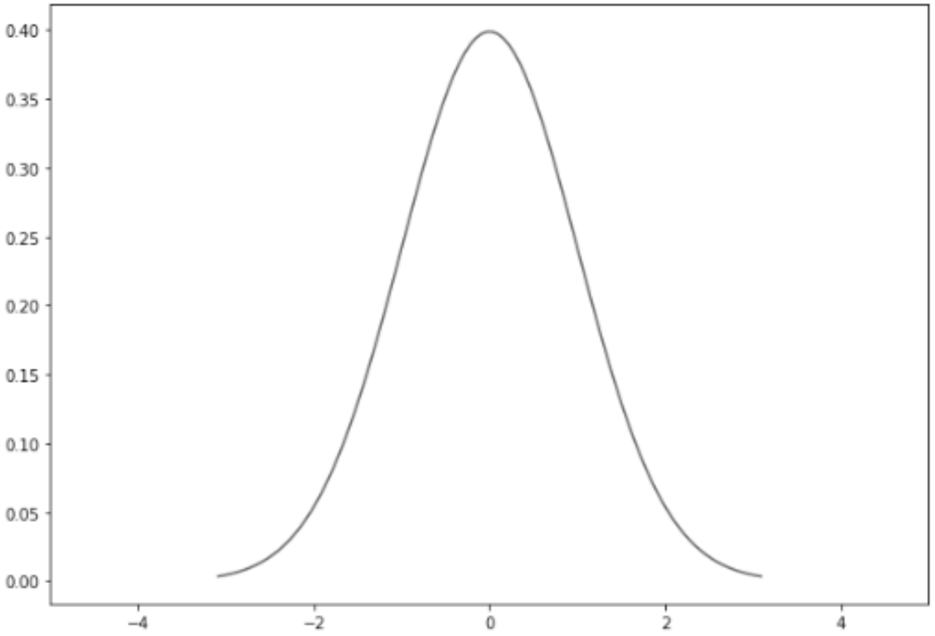
$$f_X(x) = P(X = x)$$

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

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



np. rozkład normalny



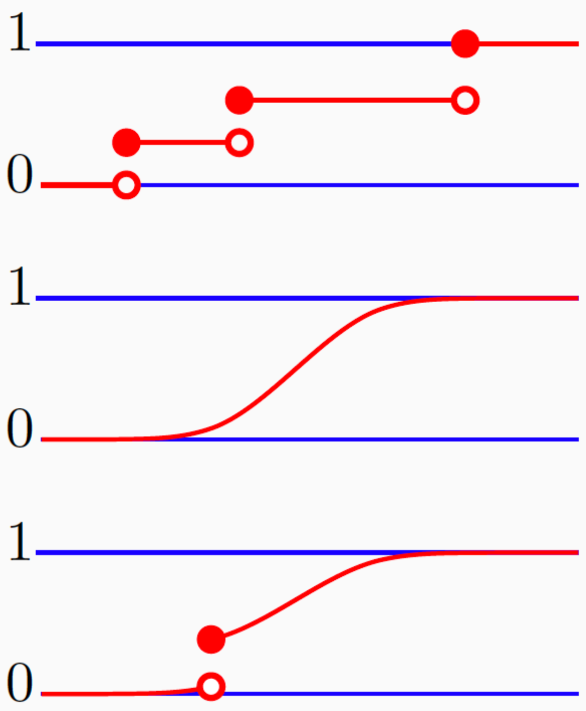
Dystrybuanta

Cumulative distribution function (CDF)

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

Każda dystrybuanta $F(x)$ jest funkcją

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



Dystrybuanta

Cumulative distribution function (CDF)

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

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

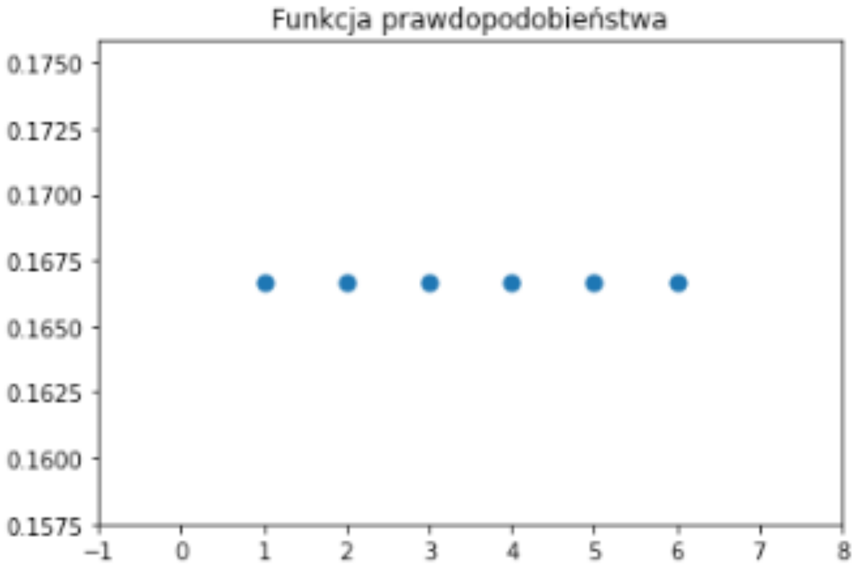
$P(X = 1) = \frac{1}{6}$ $P(X = 2) = \frac{1}{6}$ $P(X = 3) = \frac{1}{6}$ $(\ldots)$

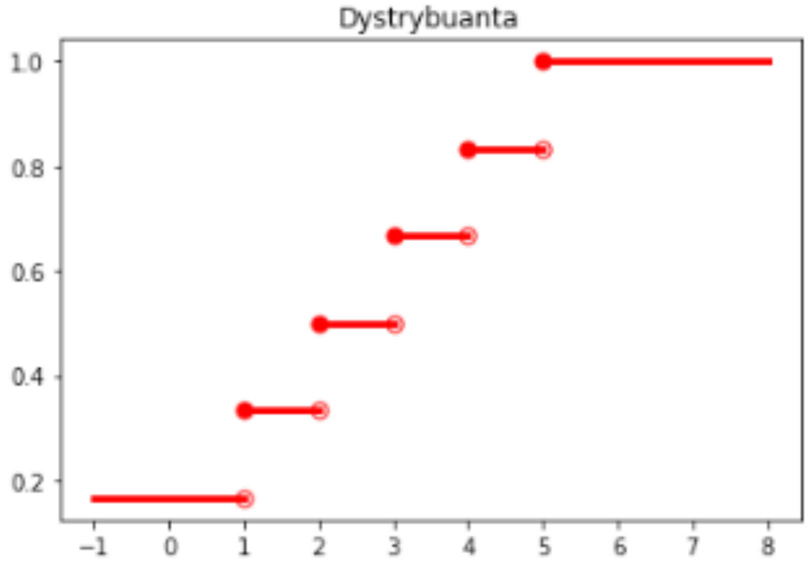
$$P(X \leq 1) = P(X = 1) = \frac{1}{6}$$

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

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

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





Parametry opisujące rozkłady

Wartość oczekiwana

Dyskretna zmienna losowa

$$\mathrm{E}[X] = \sum_{i=1}^{\infty} x_i\, p_i,$$

Ciągła zmienna losowa

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x)\, dx.$$

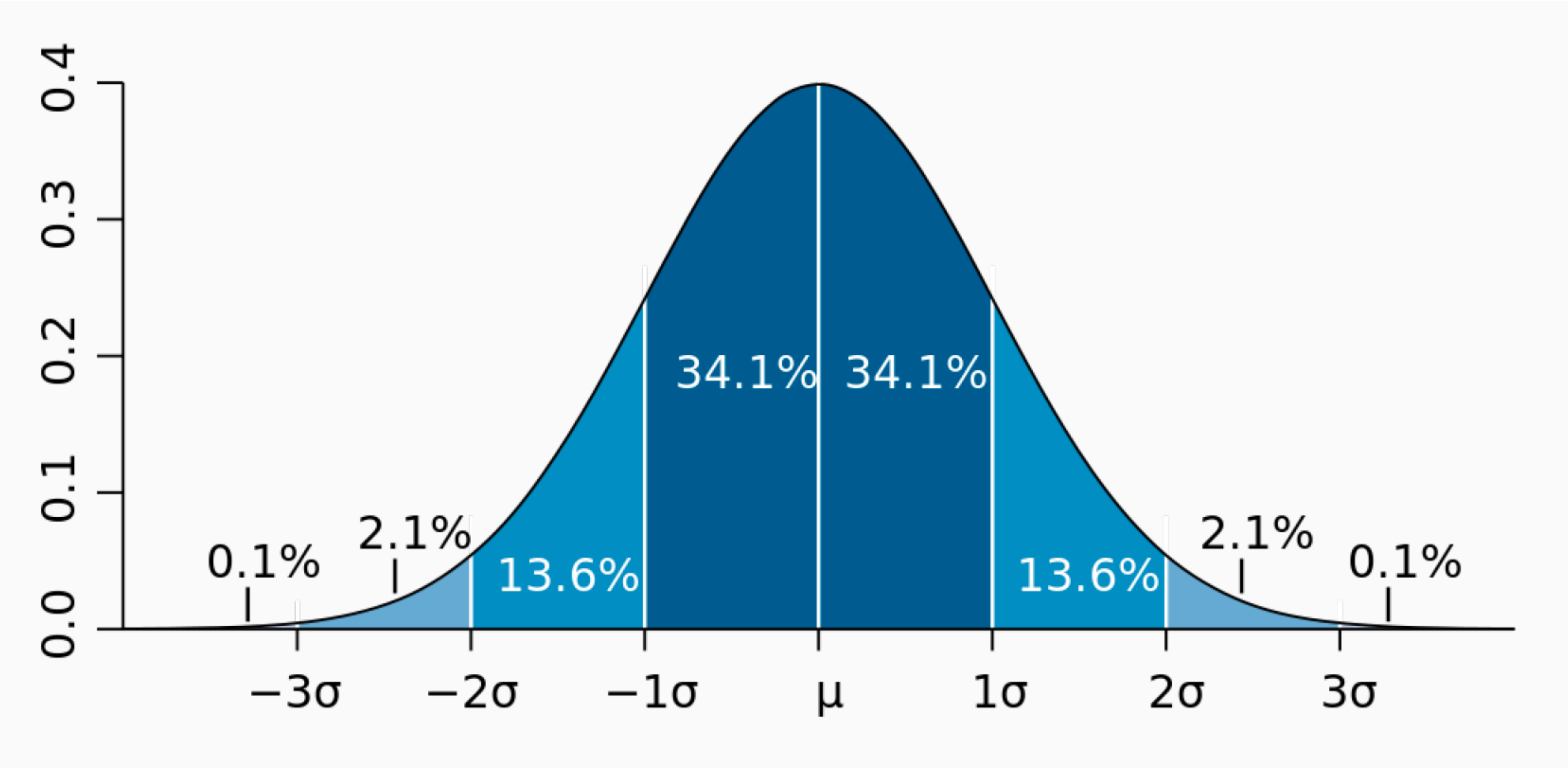
Wartość oczekiwaną nazywamy spodziewany wynik doświadczenia losowego przy założonym prawdopodobieństwie jego wystąpienia.

Wariancja

$$\begin{aligned} \operatorname{Var}(X) &= \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] \\ &= \mathrm{E}\left[X^2 - 2X\,\mathrm{E}[X] + \mathrm{E}[X]^2\right] \\ &= \mathrm{E}\left[X^2\right] - 2\,\mathrm{E}[X]\,\mathrm{E}[X] + \mathrm{E}[X]^2 \\ &= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 \end{aligned}$$

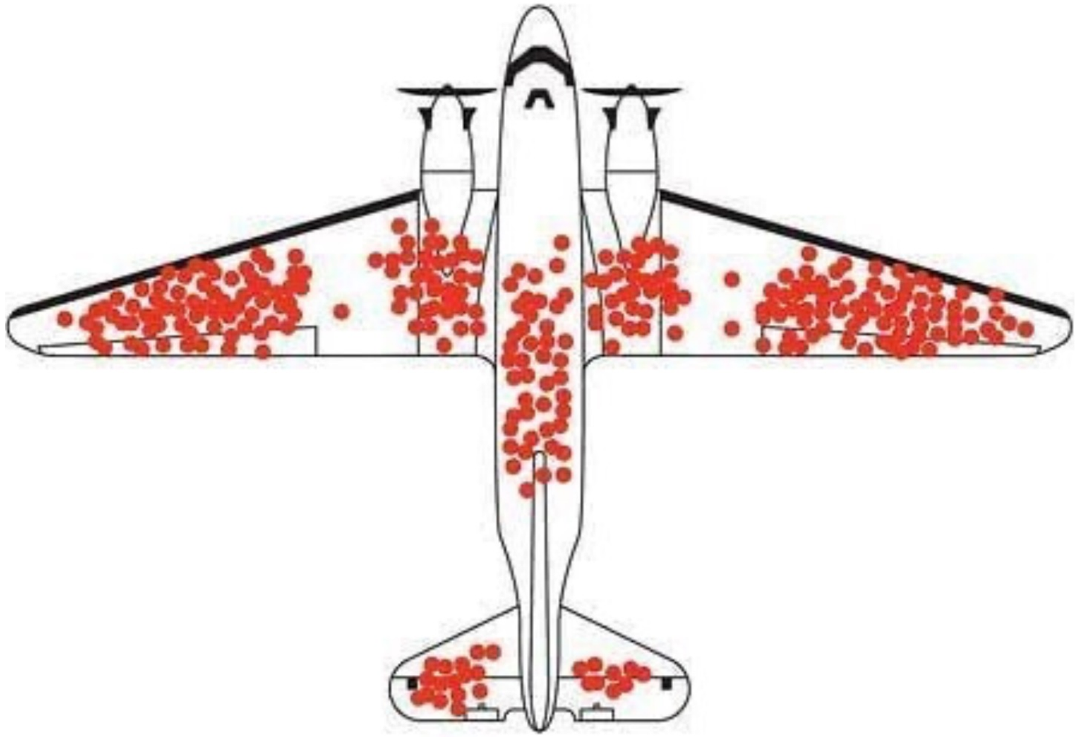
Wariancja jest podstawową miarą zmienności zmiennej losowej. Wariancja informuje o tym, jak duże jest zróżnicowanie wyników w danym zbiorze wyników (zmiennej).

Rozkład normalny – reguła trzech sigm



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

Produkcja 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

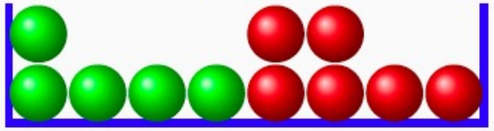
Przyjęci na studia (Univ. of California, Berkeley)

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

Przyjęci z podziałem na kierunki

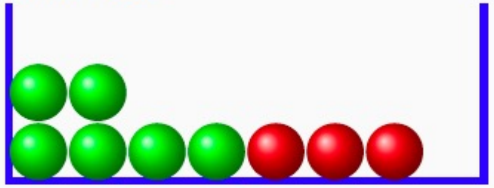
	Mężczyźni		Kobiety	
K.	Zgłoszeń	Przyjętych	Zgłoszeń	Przyjętych
A	825	(512) 62%	108	(89) 82%
B	560	(353) 63%	25	(17) 68%
C	325	(120) 37%	593	(219) 37%
D	417	(138) 33%	375	(131) 35%
E	191	(53) 28%	393	(134) 34%
F	373	(22) 6%	341	(24) 7%

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

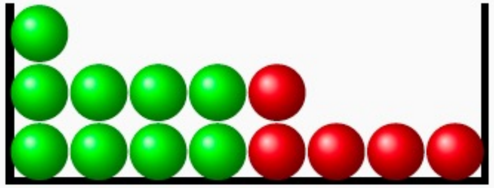


$\frac{5}{11} > \frac{3}{7}$

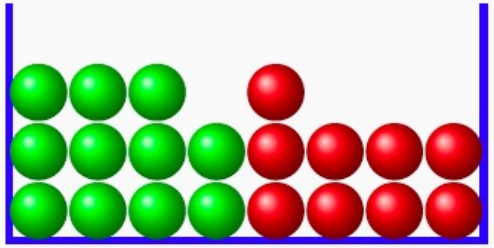
A teraz?

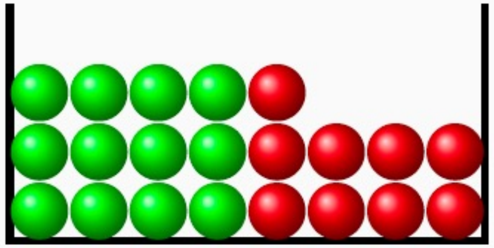


$\frac{6}{9} > \frac{9}{14}$



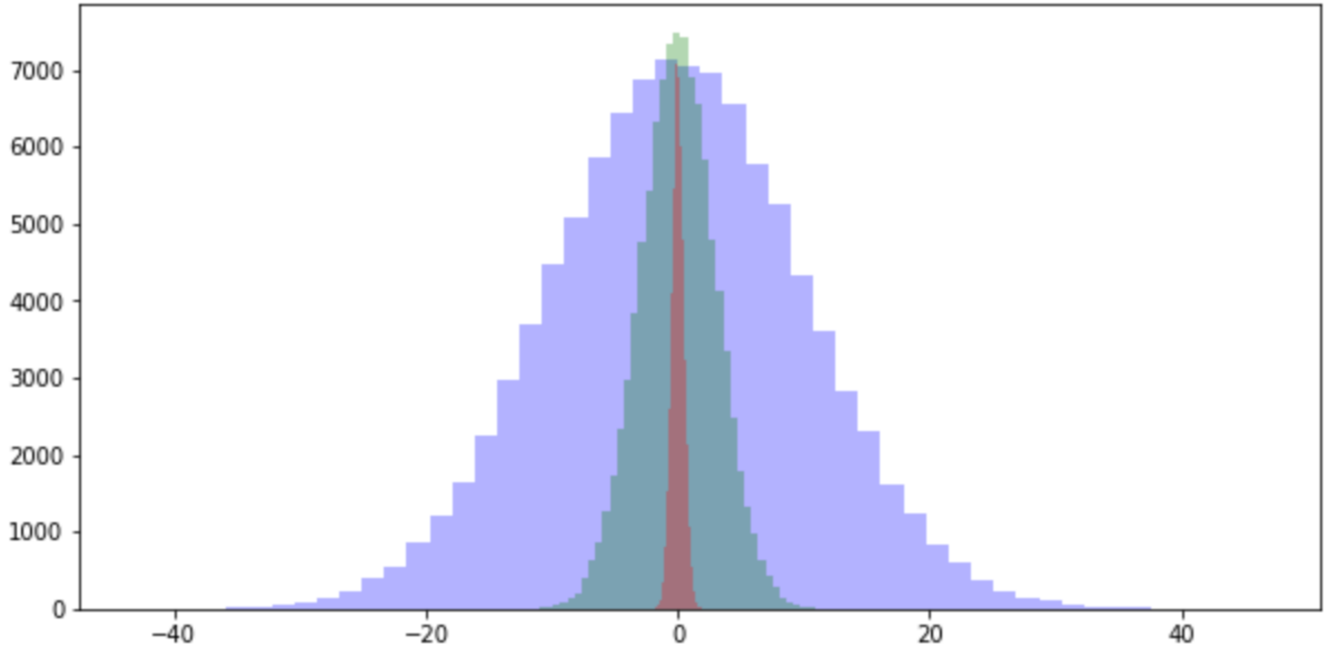
A jak połączymy zawartość koszyków?





Wariancja

```
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:

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

odchylenie standardowe:

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

kowariancja:

$$\operatorname{cov}(X, Y) = \operatorname{E}\left[(X - \operatorname{E}X) \cdot (Y - \operatorname{E}Y)\right].$$

Współczynnik korelacji Pearsona :

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

statystyka:

Wariancja populacji:

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

Wariancja próbki:

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

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