Od fragmentu po całość, czyli o sztuce uczenia się z danych

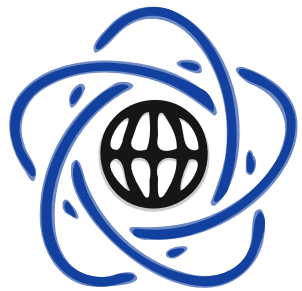
Przemysław Grzegorzewski

Wydział Matematyki i Nauk Informacyjnych

4 listopada 2017 r.

pgrzeg@mini.pw.edu.pl

http://www.ibspan.waw.pl/∼pgrzeg/



Plan wykładu

- Czym jest statystyka?
  - Trochę historii
  - Próba zdefiniowania
  - Podstawowe pojęcia
- O estymacji wskaźnika struktury
  - Estymator klasyczny i jego własności
  - Estymacja w przypadku drażliwych pytań
- O rozkładzie normalnym i rekrutach
- O analizie regresji i metodzie najmniejszych kwadratów
- Wykresy i manipulacje
- Podsumowanie

Czym jest statystyka?

Statystyka - łac. status = państwo.

Termin ten został wprowadzony przez niemieckiego uczonego Gottfrieda Achenwalla i miał oznaczać „gromadzenie, przetwarzanie i wykorzystywanie danych przez państwo" (1749 r.).

W wersji oryginalnej nazwa ta, w wyrażeniu *scienza statistica*, pojawiła się w 1579 roku.

Słowa „statystyka" w języku polskim użył po raz pierwszy S. Staszic w pracy pt. *O statystyce Polski. Krótki rzut wiadomości potrzebnych tym, którzy ten kraj chcą oswobodzić i tym którzy chcą w nim rządzić* (1809 r.).

Trochę historii

- Około 2000 p.n.e. Chiny (dynastia Sia) – spisy ludności. Za dynastii Czou (1112 – 256 p.n.e.) ustanowiono oficjalne stanowisko odpowiedzialnego za prace statystyczne „szih-su" (księgowy).
- Rzymski spis ludności został ustanowiony przez szóstego króla Rzymu Serwiusza Tuliusza (578 - 543 p.n.e.). Urzędnicy, zwani cenzorami (łac. censere - szacować) sporządzali w 5-letnich odstępach rejestr obywateli i ich własności.
- Cezar August w 5 roku p.n.e. rozszerzył spis ludności na całe Imperium Rzymskie.
- Ostatni regularny spis przeprowadzono w 74 roku n.e.
- Znane dzisiaj regularne spisy ludności zaczęły się dopiero w XVII wieku.

gromadzenie,
przetwarzanie
i wykorzystywanie danych
przez państwo

poszukiwanie
prawidłowości
występujących
w badanych zjawiskach

państwoznawstwo

arytmetyka polityczna

STATYSTYKA

- Encyclopedia Britannica (wyd. III, 1797) wzmiankuje statystykę jako *słowo wprowadzone ostatnio, aby wyrazić obraz lub zwięzły opis jakiegoś królestwa, hrabstwa lub gminy*

- 1800 – Centralny Urząd Statystyczny we Francji (pierwszy tego typu urząd na świecie)

- 1834 – Królewskie Towarzystwo Statystyczne

  fakty odnoszące się do ludzi, możliwe do przedstawienia w postaci liczb, w wystarczająco zwielokrotnionej ilości, sygnalizujące prawa ogólne

20 szczytowych odkryć od 1900 roku (Science 84, 1984)

- antybiotyki
- czaszka z Taungs
- DNA
- grupy krwi
- komputer
- lampa elektronowa
- laser
- leki przeciw chorobom umysłowym
- pestycydy
- pigułki antykoncepcyjne
- rozbicie atomu
- sieci telekomunikacyjne
- **statystyka** (test chi-kwadrat K. Pearsona)
- sztuczne nawożenie roślin
- telewizja
- teoria wielkiego wybuchu
- teoria względności Einsteina
- test IQ
- tranzystor
- tworzywa sztuczne

Problemy

Statystyka jako słowo magiczne:

- „statystyki pokazują. . . ”
- „według statystyk. . . ”
- „statystycznie. . . ”
- „badania statystyczne wykazały. . . ”

„Kłamstwo, bezczelne kłamstwo, statystyka”

„Statystyka kłamie”

Statystyka nie kłamie, ale kłamcy chętnie posługują się statystyką

„Liczby nie kłamią, ale kłamcy liczą” (C.H. Grosvenor)

Próby zdefiniowania

Statystyka – dyscyplina naukowa zajmująca się zbieraniem, prezentacją, analizą oraz interpretacją danych opisujących zjawiska masowe.

Statystyka – dyscyplina naukowa zajmująca się zarówno metodami liczbowego opisu danych (ujęcie deterministyczne), jak i metodami liczbowego wnioskowania w warunkach niepewności (ujęcie stochastyczne).

Statystyka to

- nauka
- technika
- sztuka

Czym zajmują się statystycy?

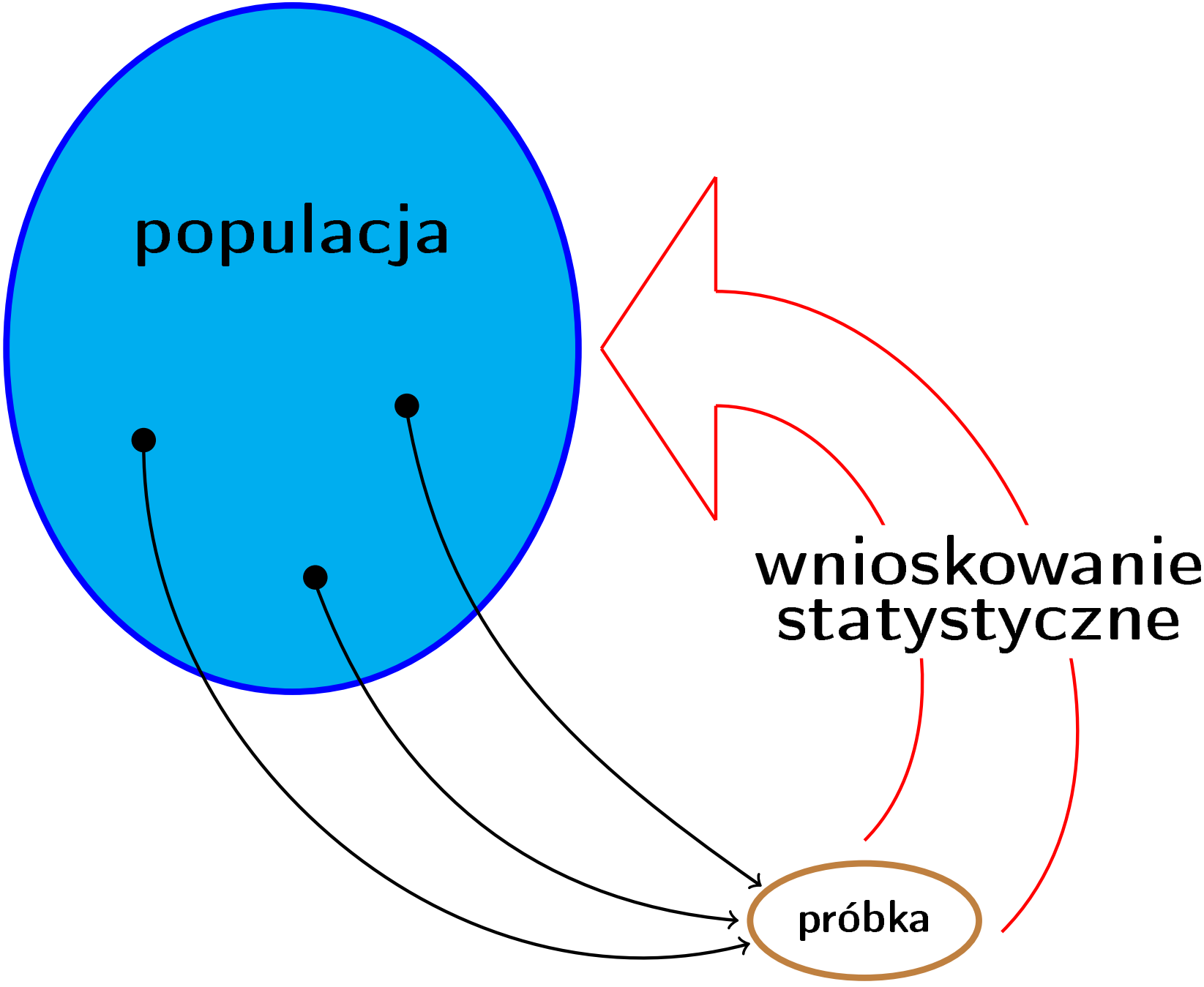
- Decydują, jakie dane są potrzebne, by odpowiedzieć na interesujące nas pytania.
- Projektują plany badania i zbierania danych.
- Przeprowadzają spisy, eksperymenty, badania, ankiety itp. mające na celu pozyskanie danych (lub szkolą w tym celu innych).
- Konstruują narzędzia do analizy danych i wspomagania wnioskowania.
- Analizują dane.
- Interpretują wyniki.
- Wyciągają wnioski z badań.

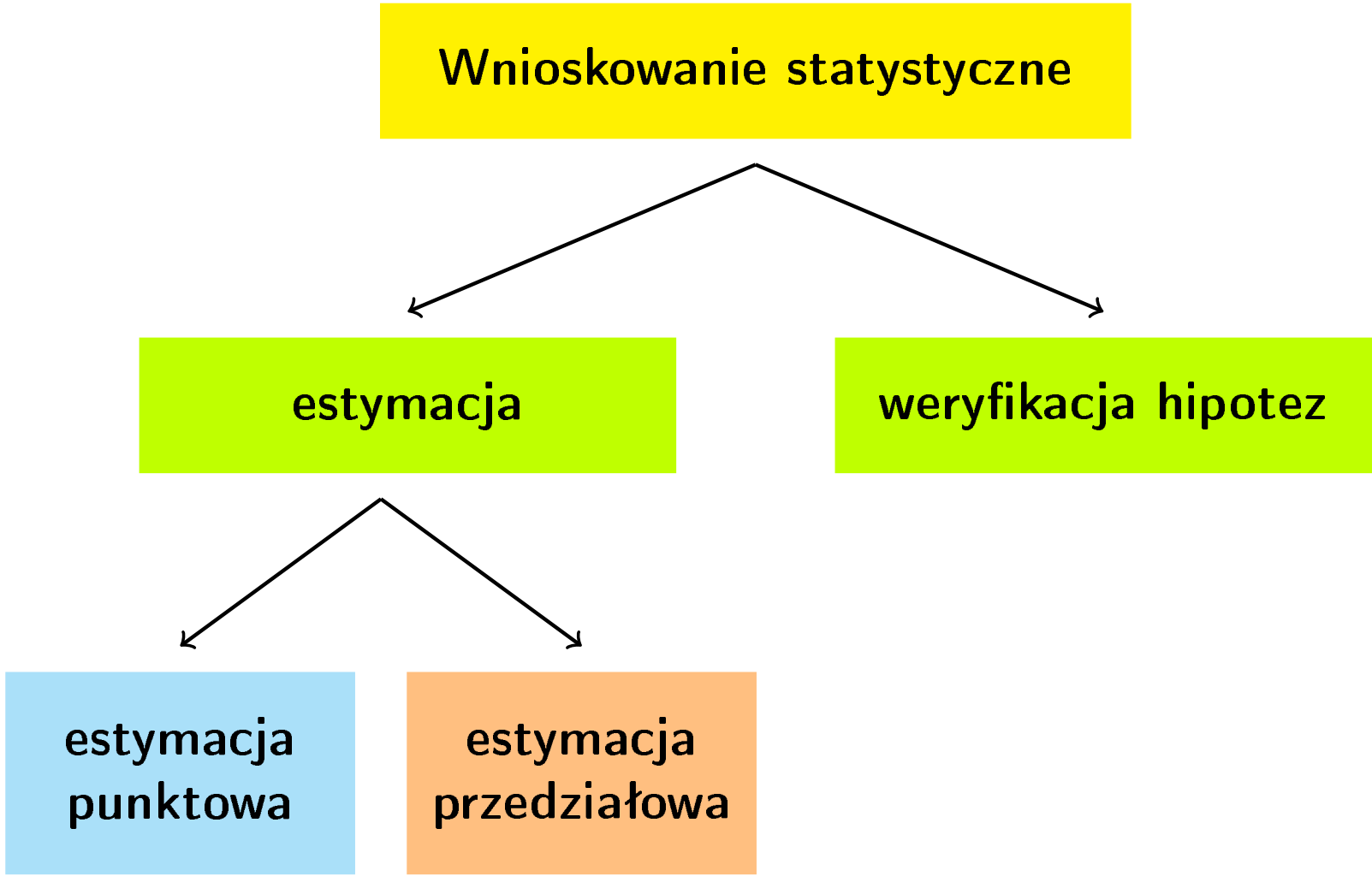
Podstawowe pojęcia statystyki

- Populacja
- Jednostka statystyczna
- Cecha
  - jakościowa (niemierzalna)
  - ilościowa (mierzalna)
    - ciągła
    - dyskretna (skokowa)
- Rozkład cechy
- Obserwacja (pomiar)
- Skala pomiarowa

	Skala pomiarowa			
	nominalna	porządkowa	przedziałowa	ilorazowa
Czy X różni się od Y?	**TAK**	**TAK**	**TAK**	**TAK**
Czy X jest większy od Y?	**NIE**	**TAK**	**TAK**	**TAK**
O ile X jest większy od Y?	**NIE**	**NIE**	**TAK**	**TAK**
Ile razy X jest większy od Y?	**NIE**	**NIE**	**NIE**	**TAK**

- Badanie
  - pełne
  - częściowe
- Próbka (próba)
  - **reprezentatywna**
  - losowa
- Dane surowe
- Wstępne przygotowanie danych (czyszczenie danych itp.)
- Statystyka opisowa
- **Wnioskowanie statystyczne**

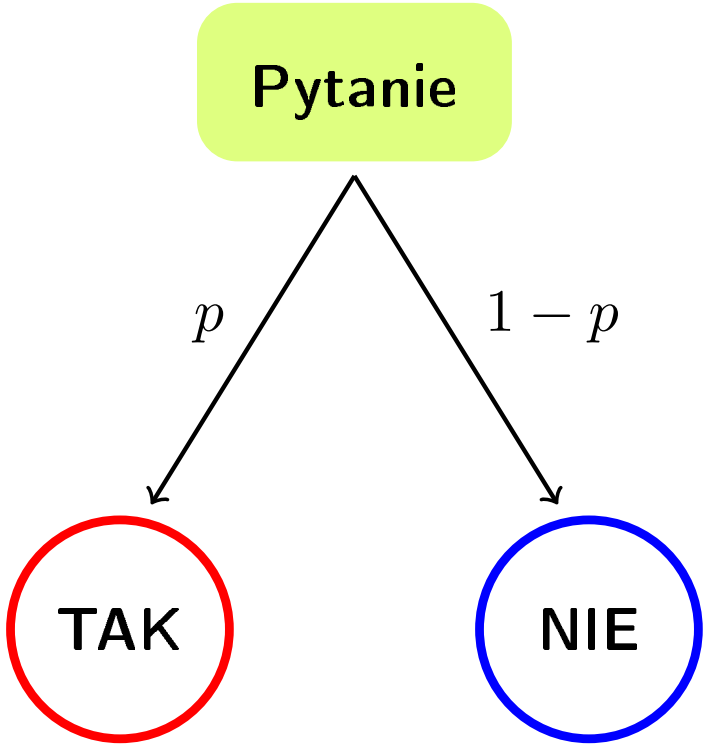

populacja
wnioskowanie
statystyczne
próbka


Wnioskowanie statystyczne
estymacja
weryfikacja hipotez
estymacja punktowa
estymacja przedziałowa

O estymacji wskaźnika struktury

(odsetka, prawdopodobieństwa sukcesu,...)

- Jaki procent dorosłych obywateli naszego kraju ma prawo jazdy?
- Jaki procent gospodarstw domowych dysponuje co najmniej jednym samochodem?
- Jaki odsetek dzieci umie pływać?
- Jak dużym poparciem cieszy się kandydat ... ?
- ...



Jak oszacować wartość p?

n – liczba pytanych osób

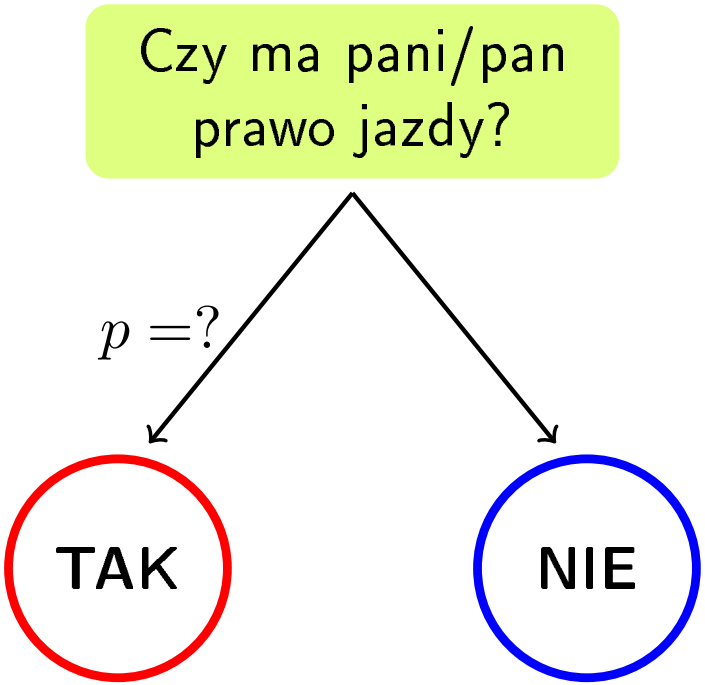
k – liczba odpowiedzi TAK

$$\widehat{p} = \frac{k}{n}(\cdot 100\%)$$

$\widehat{p}$ – estymator wartości p

Przykład

Spośród 800 osób zapytanych: „Czy ma pani/pan prawo jazdy?”, 432 odpowiedziały twierdząco.
Oszacować odsetek osób posiadających prawo jazdy.



$$
\begin{aligned}
n &= 800 \\
k &= 432 \\
\\
\widehat{p} &= \frac{k}{n} \\
&= \frac{432}{800} = 0.54
\end{aligned}
$$

Drażliwe pytania

Jak szacować odsetek w przypadku tzw. drażliwych pytań?

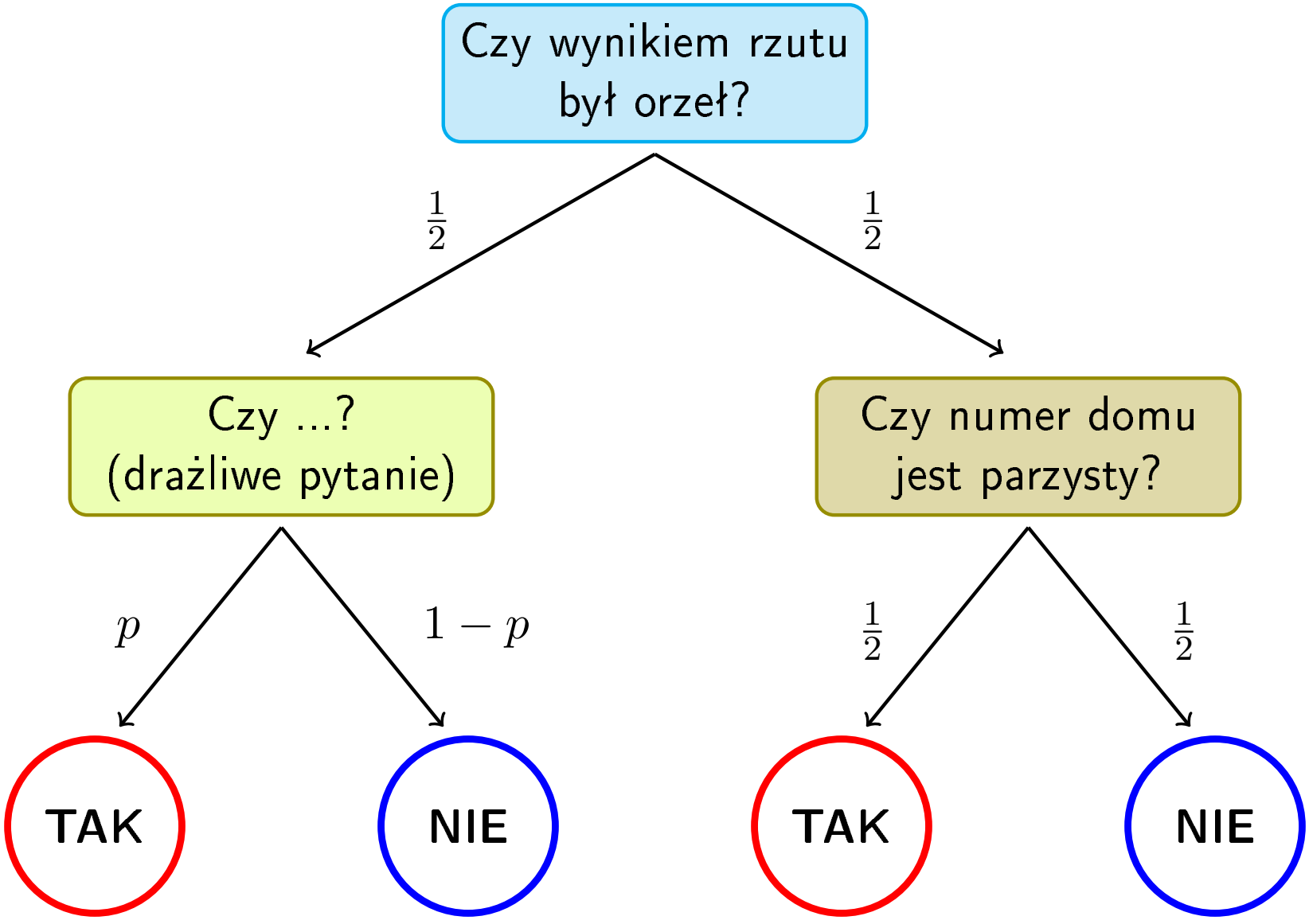
Przykładowo:

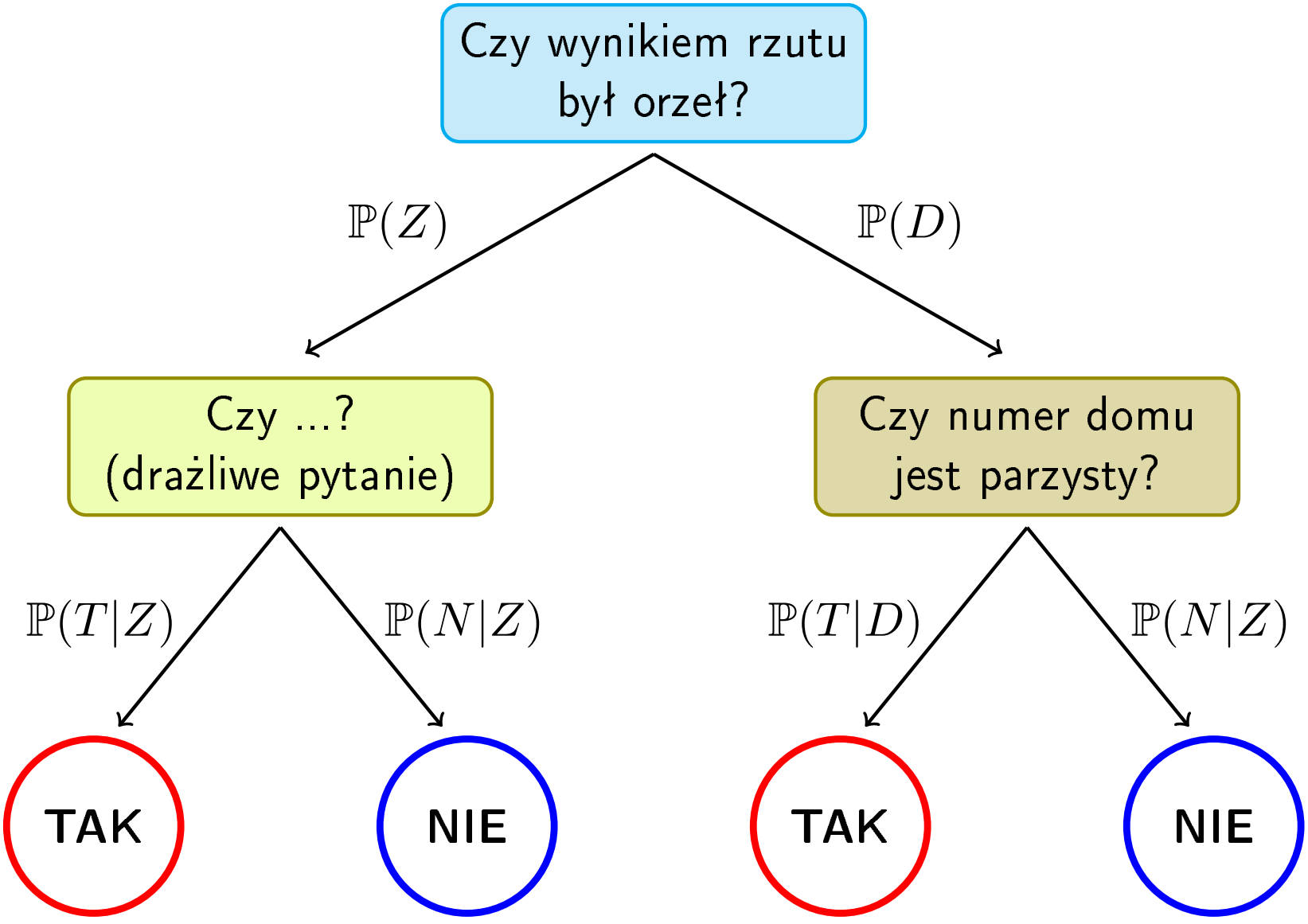
- Czy brał/dawał pan łapówkę?
- Czy zatrudnia pani/pan pracowników „na czarno”?
- Czy płaci pani/pan podatki?
- Czy zażywasz narkotyki?
- Czy dokonała pani aborcji?
- ...

Ankieter, zwracając się do respondenta z drażliwym pytaniem, mówi:

Zdaję sobie sprawę, że odpowiadanie na to pytanie może być krępująca. Postąpmy zatem następująco:

- Proszę rzucić monetą, ale nie pokazywać wyniku rzutu.
- Jeśli wypadnie orzeł , proszę odpowiedzieć uczciwie tak/nie na zadane pytanie.
- Natomiast, jeśli wypadnie reszka, proszę odpowiedzieć tak/nie na pytanie: „czy numer domu, w którym pani/pan mieszka jest parzysty".

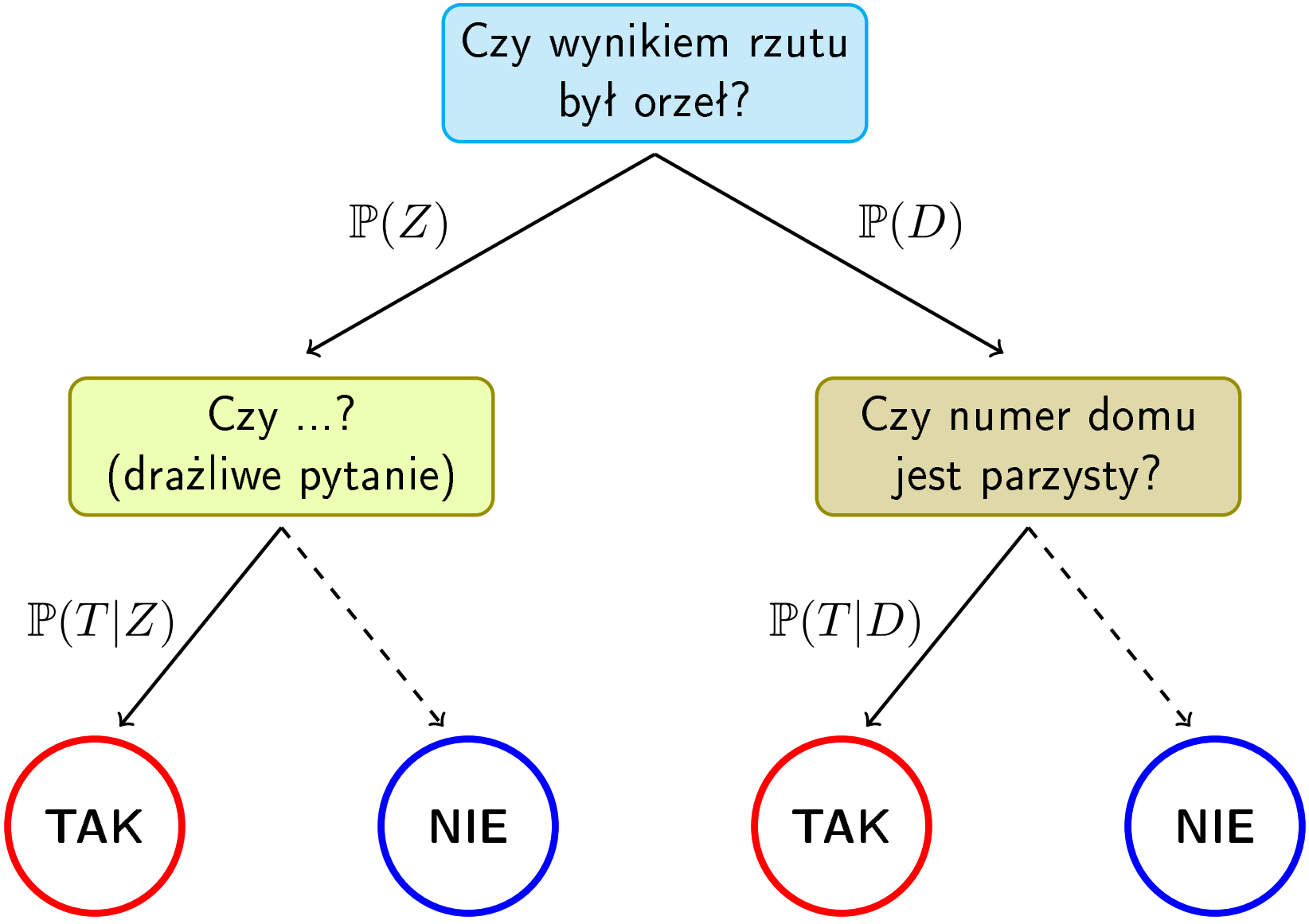

Czy wynikiem rzutu był orzeł?
$\frac{1}{2}$
$\frac{1}{2}$
Czy ...? (drażliwe pytanie)
Czy numer domu jest parzysty?
p
$1-p$
$\frac{1}{2}$
$\frac{1}{2}$
TAK
NIE
TAK
NIE

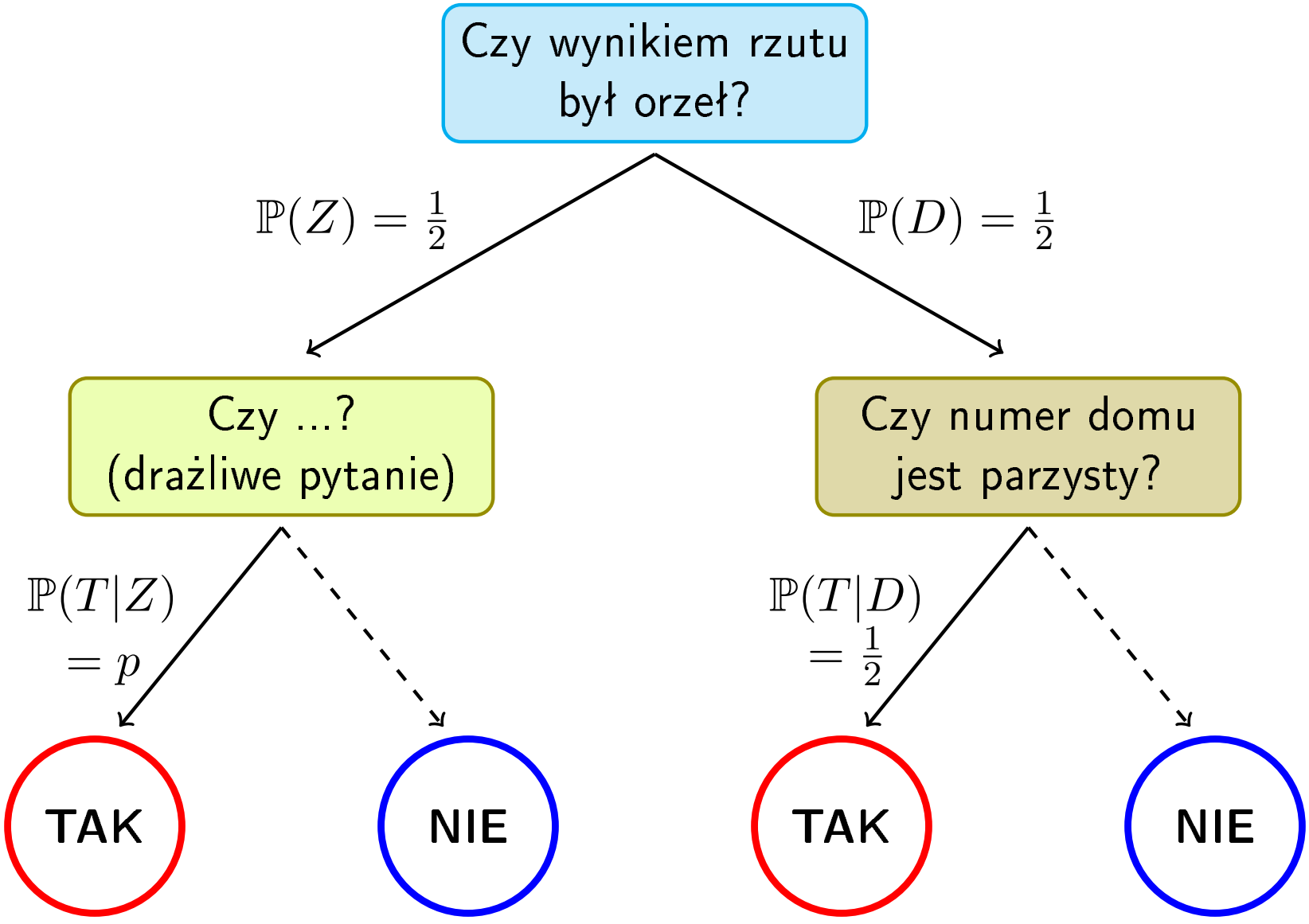

Czy wynikiem rzutu był orzeł?
$\mathbb{P}(Z)$
$\mathbb{P}(D)$
Czy ...? (drażliwe pytanie)
Czy numer domu jest parzysty?
$\mathbb{P}(T|Z)$
$\mathbb{P}(N|Z)$
$\mathbb{P}(T|D)$
$\mathbb{P}(N|Z)$
TAK
NIE
TAK
NIE

Twierdzenie (o prawdopodobieństwie całkowitym)

Jeżeli Z i D są zdarzeniami wykluczającymi się, to prawdopodobieństwo zajścia zdarzenia T jest dane wzorem:

$$\mathbb{P}(T) = \mathbb{P}(T|Z) \cdot \mathbb{P}(Z) + \mathbb{P}(T|D) \cdot \mathbb{P}(D).$$


Czy wynikiem rzutu był orzeł?
$\mathbb{P}(Z)$
$\mathbb{P}(D)$
Czy ...? (drażliwe pytanie)
Czy numer domu jest parzysty?
$\mathbb{P}(T|Z)$
$\mathbb{P}(T|D)$
TAK
NIE
TAK
NIE


Czy wynikiem rzutu był orzeł?
$\mathbb{P}(Z) = \frac{1}{2}$
$\mathbb{P}(D) = \frac{1}{2}$
Czy ...? (drażliwe pytanie)
Czy numer domu jest parzysty?
$\mathbb{P}(T|Z) = p$
$\mathbb{P}(T|D) = \frac{1}{2}$
TAK
NIE
TAK
NIE

W naszym przypadku otrzymujemy

$$\begin{aligned}\mathbb{P}(T) &= \underbrace{\mathbb{P}(T|Z)}_{p} \cdot \underbrace{\mathbb{P}(Z)}_{\frac{1}{2}} + \underbrace{\mathbb{P}(T|D)}_{\frac{1}{2}} \cdot \underbrace{\mathbb{P}(D)}_{\frac{1}{2}} \\ &= p \cdot \frac{1}{2} + \frac{1}{4}\end{aligned}$$

Prawdopodobieństwo $\mathbb{P}(T)$ estymujemy na podstawie wyników badania, tzn. że w przypadku uzyskania k odpowiedzi TAK od n ankietowanych osób mamy

$$\widehat{\mathbb{P}(T)} = \frac{k}{n}.$$

Zatem

$$\frac{k}{n} \approx p \cdot \frac{1}{2} + \frac{1}{4}$$

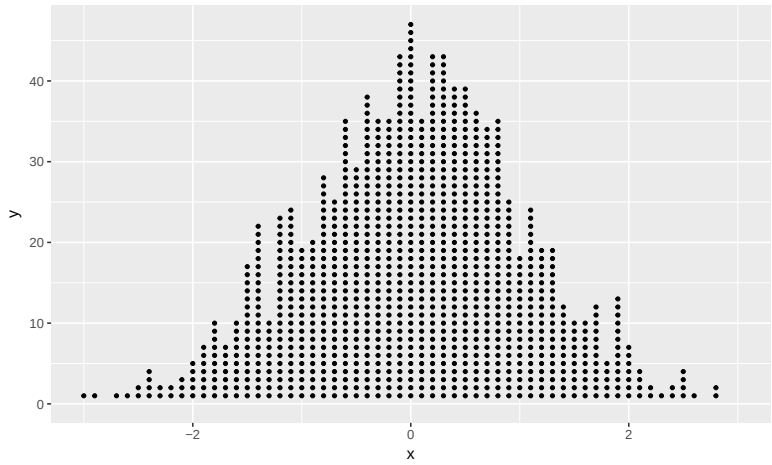
$$\frac{k}{n} \approx p \cdot \frac{1}{2} + \frac{1}{4} \qquad \Longrightarrow \qquad p \approx 2 \cdot \frac{k}{n} - \frac{1}{2}$$

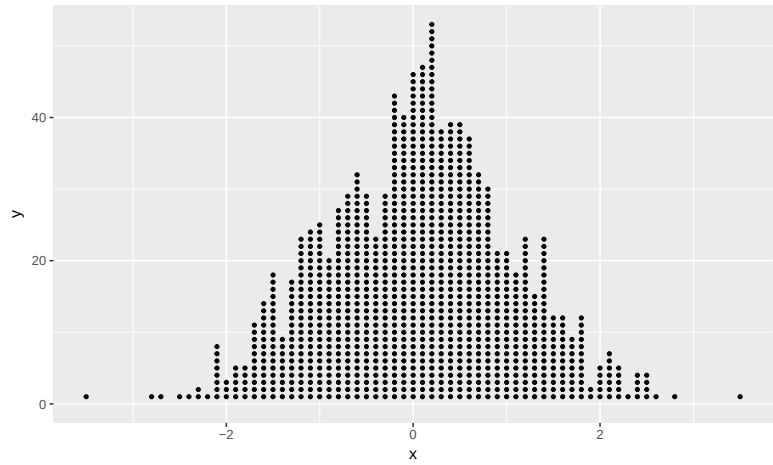
Stąd wzór na estymator odsetka, przydatny w przypadku drażliwych pytań, ma postać

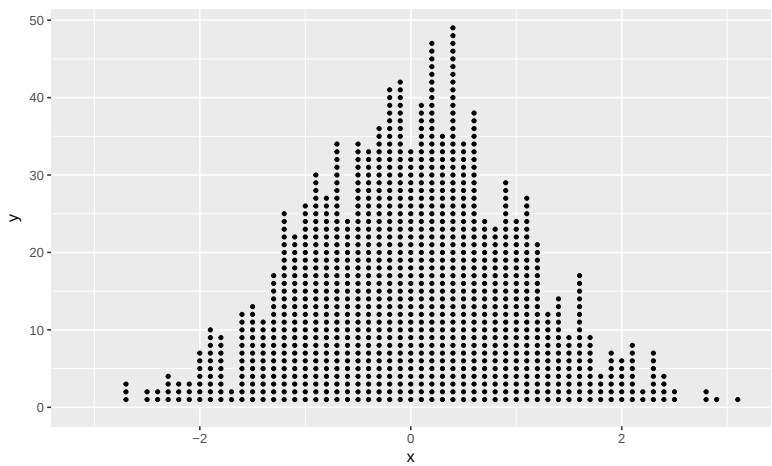
$$\widehat{p}_d = 2 \cdot \frac{k}{n} - \frac{1}{2}$$

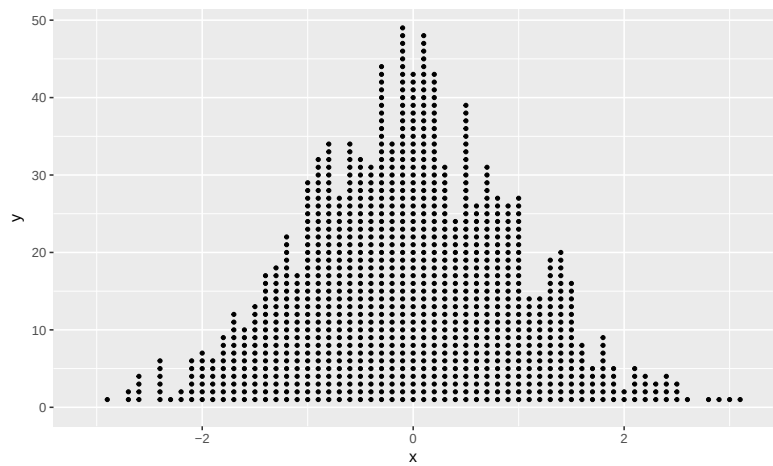
Uwaga! Ponieważ musi zachodzić $0 < \widehat{p}_d < 1$, zatem powyższy wzór można stosować gdy $\frac{n}{4} < k < \frac{3n}{4}$.

O rozkładzie normalnym i rekrutach

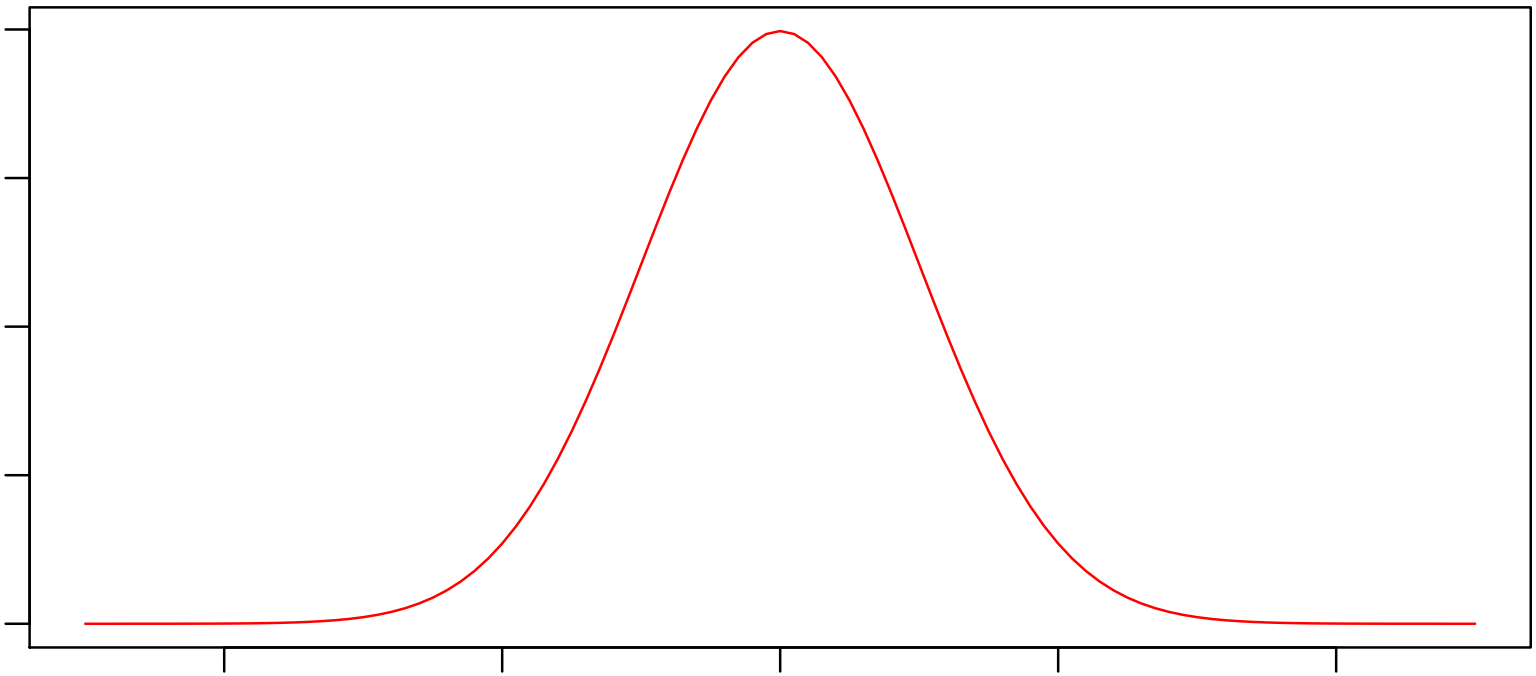

y
x
0
10
20
30
40
−2
0
2


y
x
0
20
40
−2
0
2


y
x
0
10
20
30
40
50
−2
0
2


y
x
0
10
20
30
40
50
−2
0
2

Rozkład normalny (gaussowski)

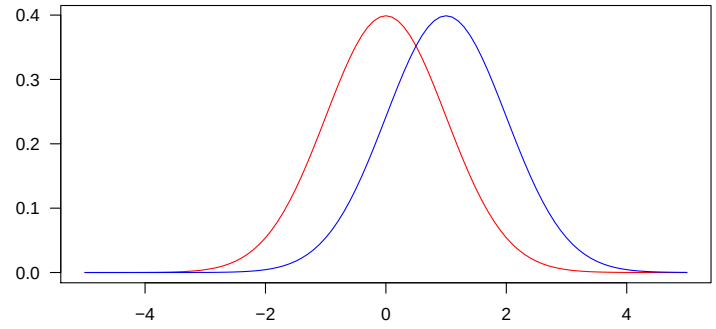


$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

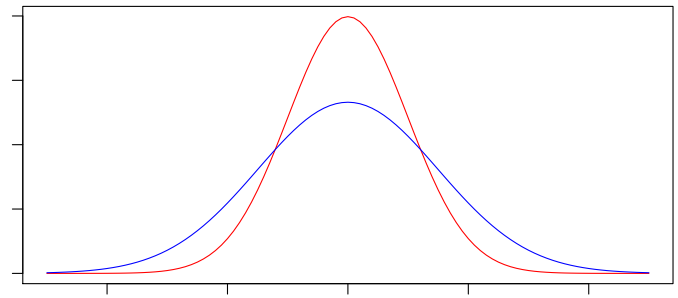
gdzie

$\mu \in \mathbb{R}$, $\sigma > 0$,

$e = 2,7182818284 59...$



$\mu_C < \mu_N$, $\sigma_C = \sigma_N$

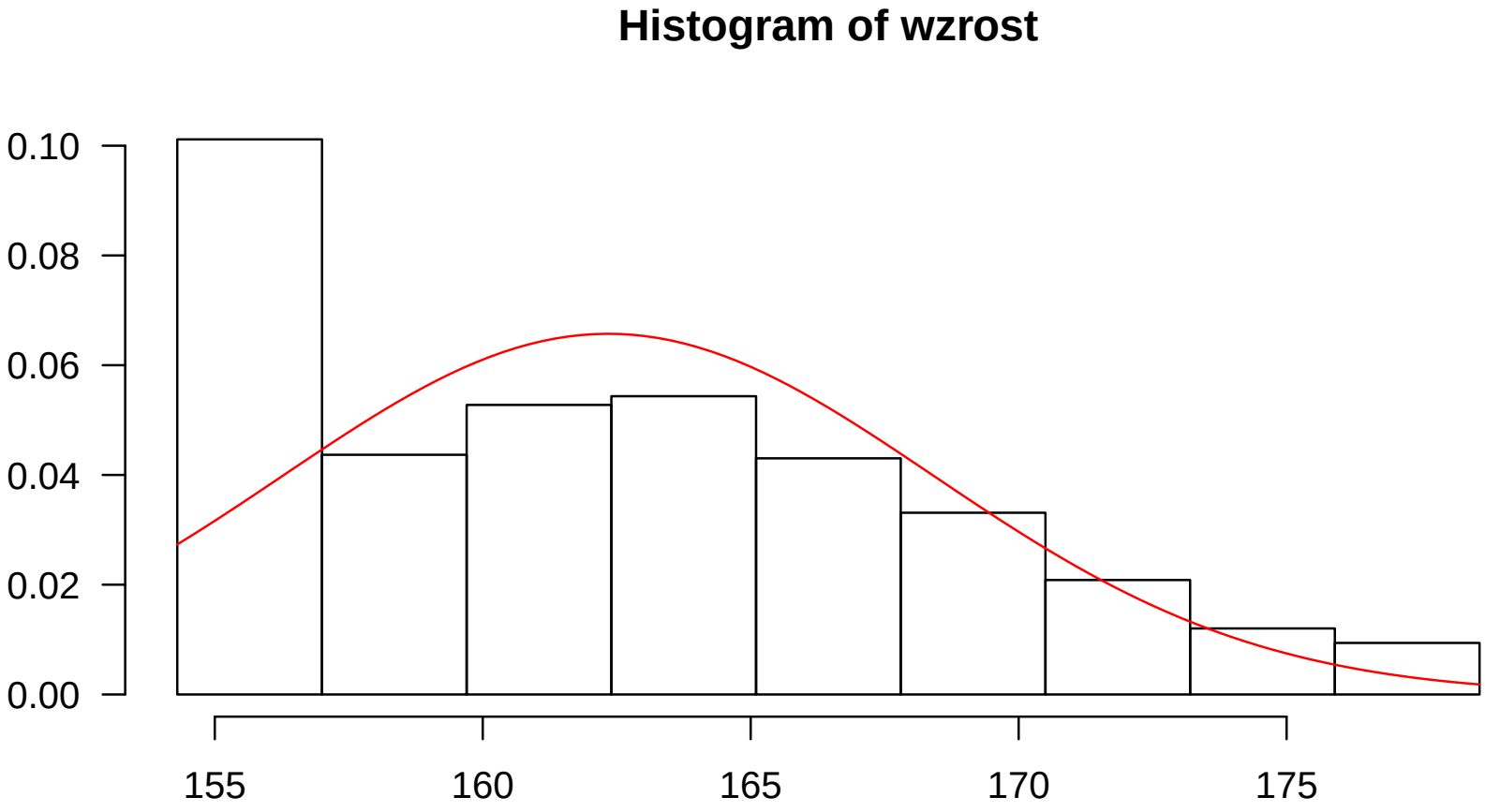


$\mu_C = \mu_N$, $\sigma_C < \sigma_N$

Lambert Adolphe Quetelet
(1796 – 1874)

Sformułował hipotezę, że krzywą rozkładu normalnego można stosować w naukach o charakterze biologiczno-społecznym z takim samym powodzeniem jak w astronomii. W szczególności, wykorzystywał ją do analizy i prognozowania przestępstw kryminalnych.

W 1844 r. określił zasięg uchylania się od poboru do wojska we Francji (modelując wzrostu mężczyzn za pomocą rozkładu normalnego).


Histogram of wzrost
0.10
0.08
0.06
0.04
0.02
0.00
155
160
165
170
175

O analizie regresji

i metodzie najmniejszych kwadratów

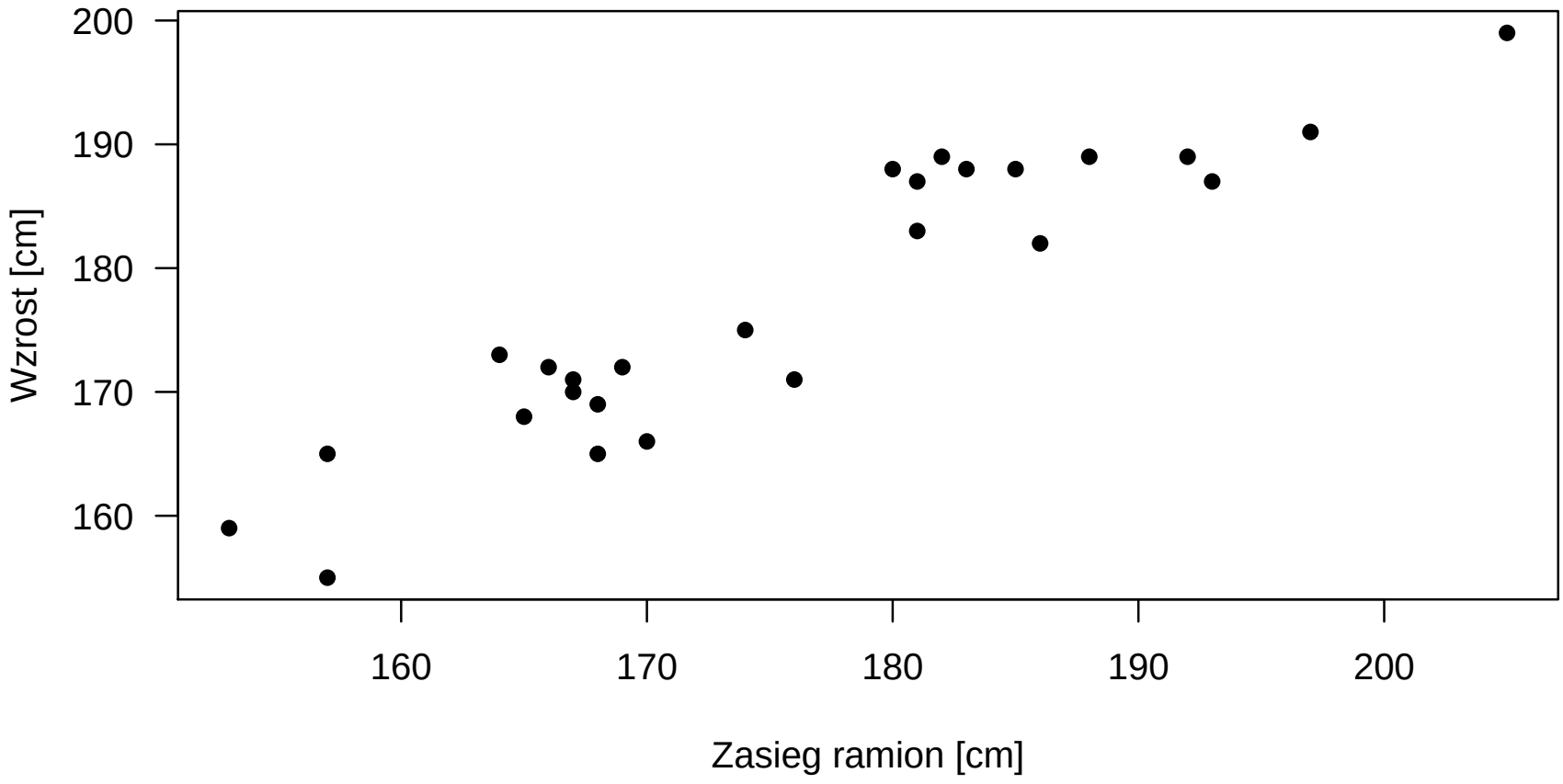
Przykład

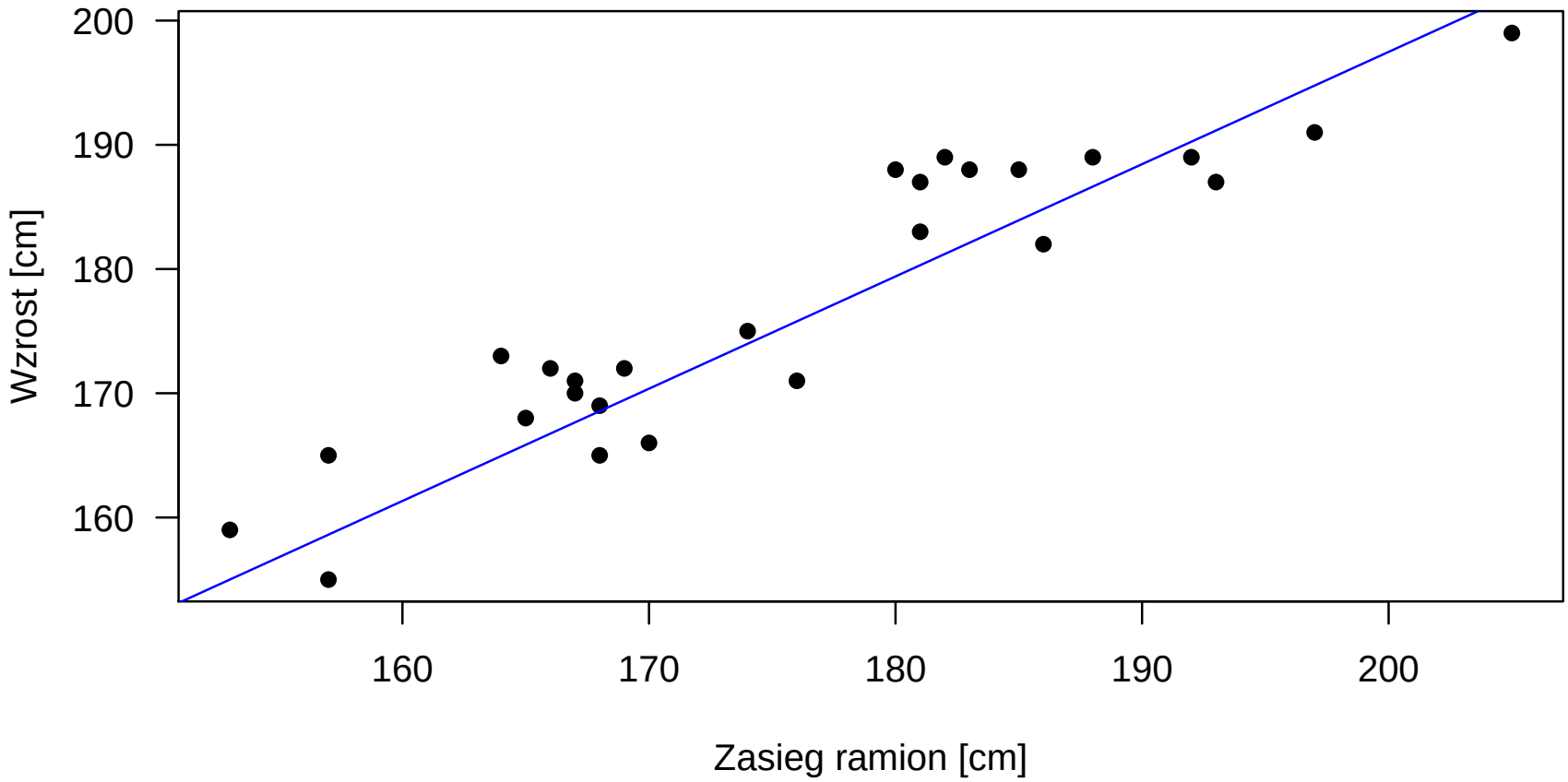
Zasięg ramion i wzrost losowej grupy studentów:

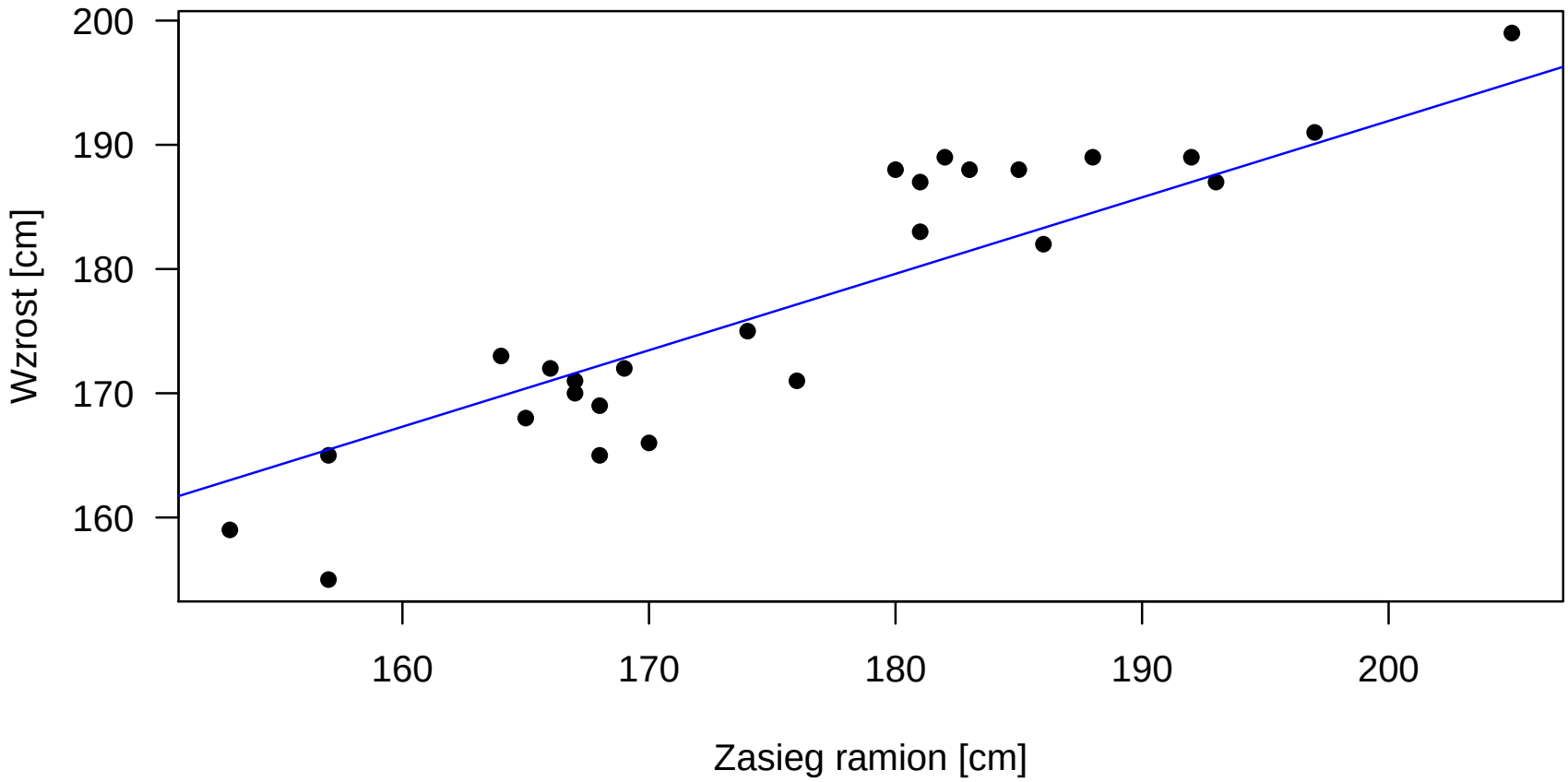
Student	Zasięg ramion	Wzrost	Student	Zasięg ramion	Wzrost
1	153	159	15	180	188
2	157	155	16	181	183
3	157	165	17	181	187
4	164	173	18	182	189
5	165	168	19	183	188
6	166	172	20	185	188
7	167	170	21	186	182
8	167	171	22	188	189
9	168	169	23	192	189
10	168	165	24	193	187
11	169	172	25	195	183
12	170	166	26	197	191
13	174	175	27	205	199
14	176	171			

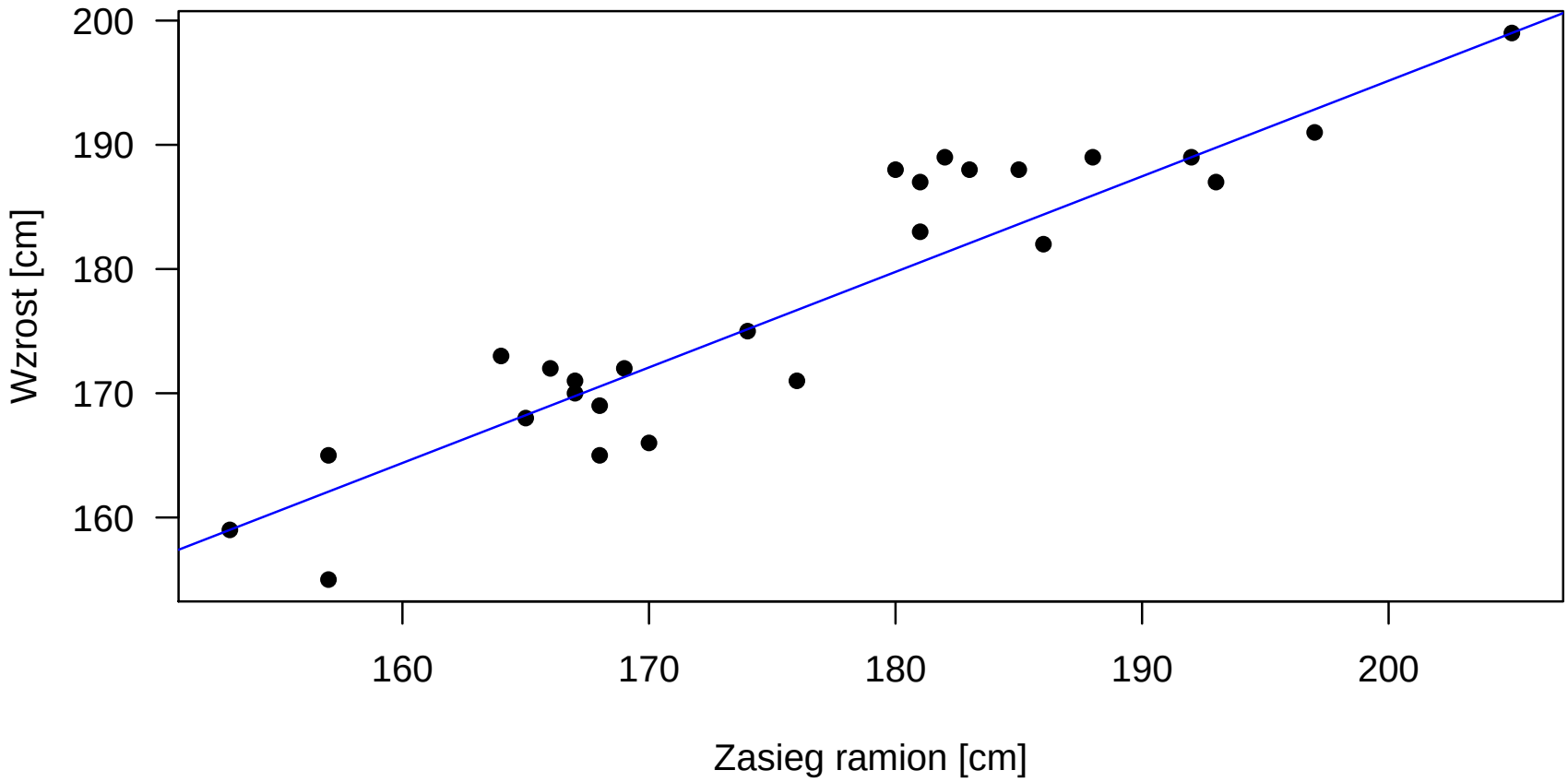
1) **Czy istnieje zależność między wzrostem i zasięgiem ramion?**

2) **Jesli tak, to jak znaleźć funkcję opisującą tę zależność?**


Wzrost [cm]
200
190
180
170
160
160
170
180
190
200
Zasieg ramion [cm]


Wzrost [cm]
200
190
180
170
160
160
170
180
190
200
Zasieg ramion [cm]


Wzrost [cm]
200
190
180
170
160
160
170
180
190
200
Zasieg ramion [cm]


Wzrost [cm]
200
190
180
170
160
160
170
180
190
200
Zasieg ramion [cm]

Model analizy regresji:

$$Y = f(X) + \epsilon,$$

gdzie
Y – zmienna objaśniana (wzrost)
X – zmienna objasniająca (zasięg ramion)
f – poszukiwana funkcja regresji
ϵ – składnik losowy (błąd).

Model liniowy:

$$Y = a \cdot X + b + \epsilon,$$

gdzie a i b – poszukiwane stałe (współczynniki prostej regresji)

Jak wyznaczyć współczynniki optymalnej prostej regresji?

- Czym dysponujemy?

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

- Jakie przyjąć kryterium optymalności?

Minimalizacja błędu średniokwadratowego

Metoda najmniejszych kwadratów

$$\underbrace{e_1^2 + e_2^2 \ldots + e_n^2}_{\text{błąd średniokwadratowy}} \longrightarrow \min$$

gdzie

$$\begin{aligned} e_1 &= y_1 - (ax_1 + b) \\ e_2 &= y_2 - (ax_2 + b) \\ &\ldots \\ e_n &= y_n - (ax_n + b) \end{aligned}$$

Skąd się wzięła nazwa metody?

Rozwiązanie zadania:

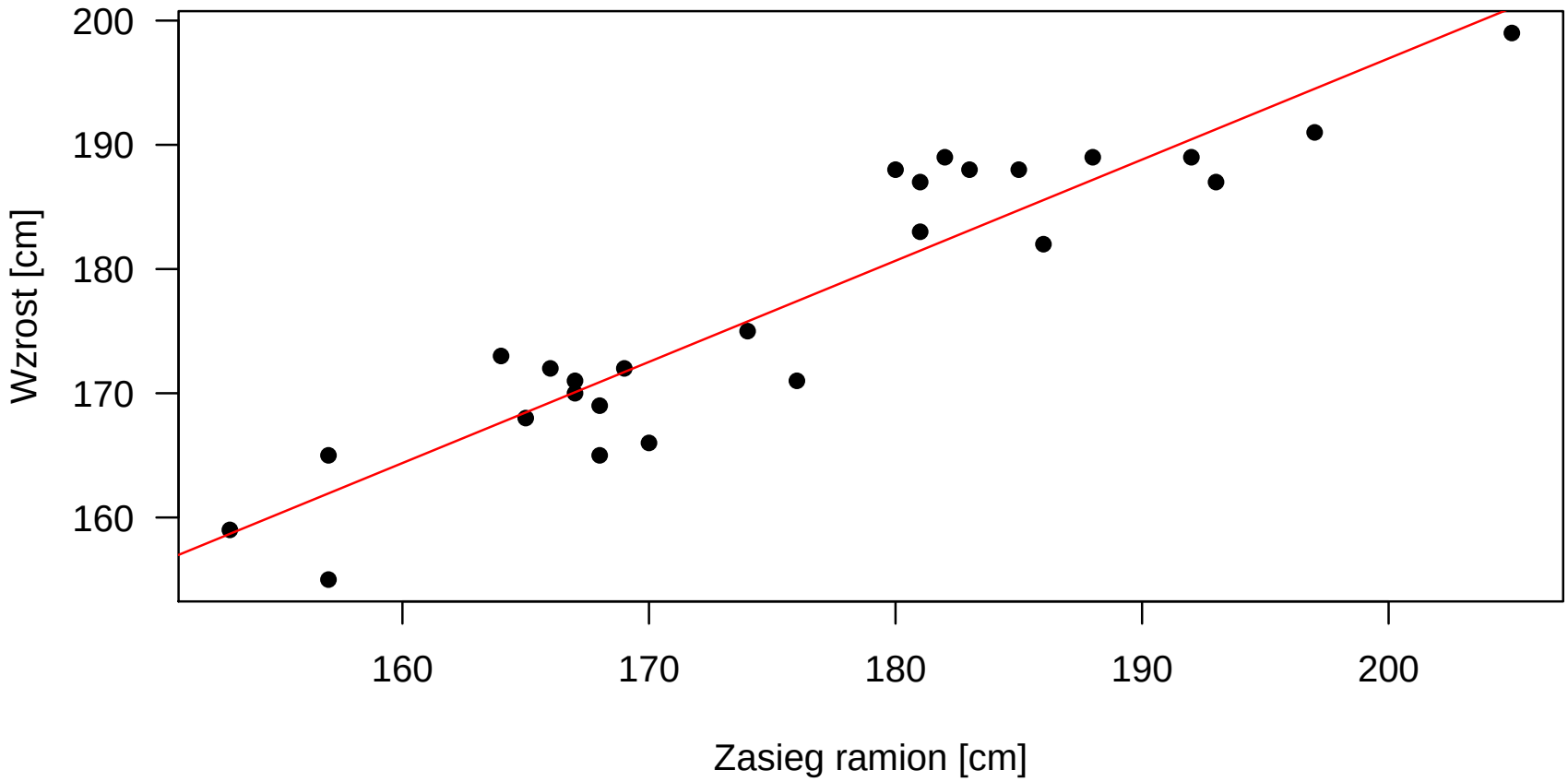
$$\begin{aligned} a &= \frac{(x_1y_1 + x_2y_2 + \ldots + x_ny_n) - n(\overline{x})(\overline{y})}{(x_1^2 + x_2^2 + \ldots + x_n^2) - n(\overline{x})^2} \\ b &= \overline{y} - a\overline{x}, \end{aligned}$$

gdzie

$$\begin{aligned} \overline{x} &= \frac{1}{n}\big(x_1 + x_2 + \ldots + x_n\big) \\ \overline{y} &= \frac{1}{n}\big(y_1 + y_2 + \ldots + y_n\big) \end{aligned}$$

W naszym przykładzie otrzymamy:

$$\begin{aligned} a &= 0.8143 \\ b &= 34.0938 \end{aligned}$$


Wzrost [cm]
200
190
180
170
160
160
170
180
190
200
Zasieg ramion [cm]

Przykładowe zastosowanie analizy regresji:

- Jakiego wzrostu (mniej więcej) będzie osoba o zasiegu ramion równym 179 cm?

$$y = 0.8143 \cdot 179 + 34.0938 = 179.04 \text{ cm}$$

- Jaki zasięg ramion (mniej więcej) będzie miała osoba o wzroście 200 cm?

$$0.8143 \cdot x + 34.0938 = 200$$

$$\Downarrow$$

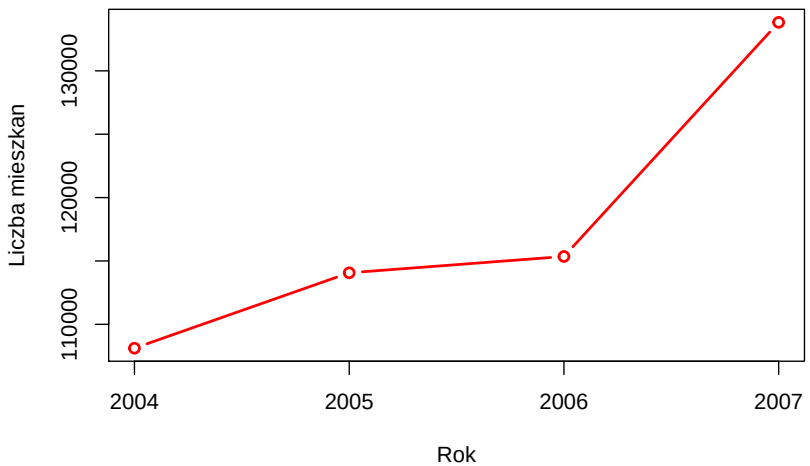
$$x = \frac{200 - 34.0938}{0.8143} = 203.74 \text{ cm}$$

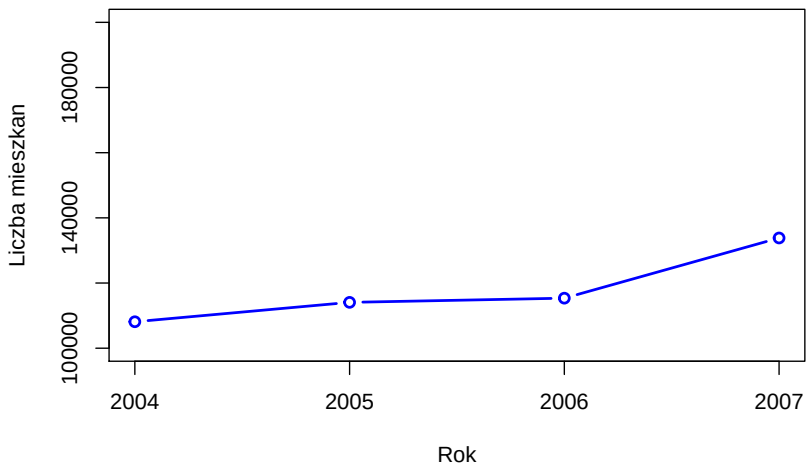
Wykresy i manipulacje

Przykład

Liczba mieszkań oddanych do użytkowania w latach 2004 – 2007:

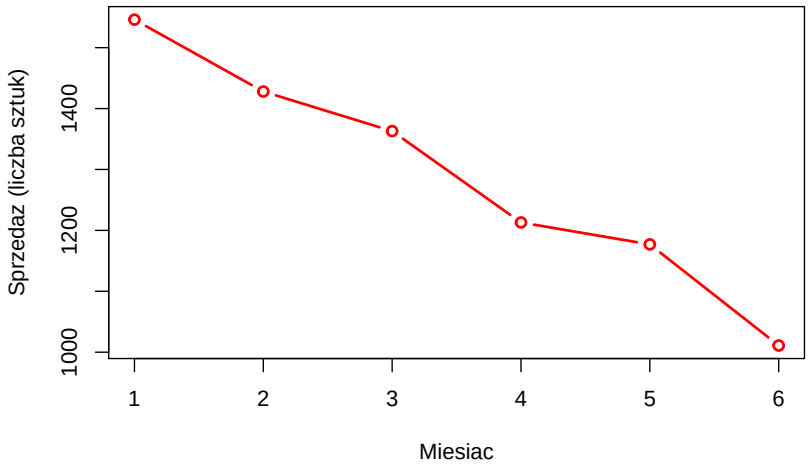
Rok	Liczba mieszkań
2004	108117
2005	114066
2006	115353
2007	133826

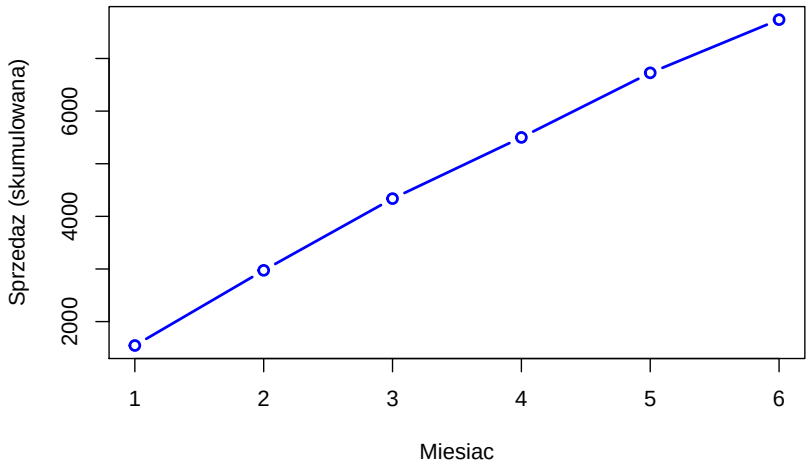




Przykład

Wyniki sprzedaży:





Podsumowanie

Bądź zawsze krytyczny wobec danych i sposobów ich prezentacji – bez względu na ich genezę oraz to, gdzie są publikowane!

Wykres może być wart tysiąca słów ... ale też od wykresów mogą zaczynać się tzw. „statystyczne kłamstwa".

Nawet najbardziej wyrafinowane metody statystyczne mogą okazać się niewiele warte, jeśli dane, którymi dysponujemy, będą wątpliwej jakości.

Podsumowanie