SZACOWANIE ROZMIARU BŁĘDU WYNIKAJĄCEGO Z OBECNOŚCI BRAKÓW DANYCH

2013-12-15

Piotr Zielonka

Braki danych

- Braki udziału
- Braki odpowiedzi
- Odpowiedzi beztreściowe

Wszystkie te zjawiska będę nazywał brakami danych.

Odsetek braków danych

$$\frac{\text{Liczba braków udziału} + \text{Liczba odmów odpowiedzi} + \text{Liczba odpowiedzi beztreściowych}}{\text{Liczba jednostek wylosowanych do próby}} = \text{Odsetek braków danych}$$

Liczby jednostek wylosowanych do próby nie należy mylić z liczebnością przebadanej próby.

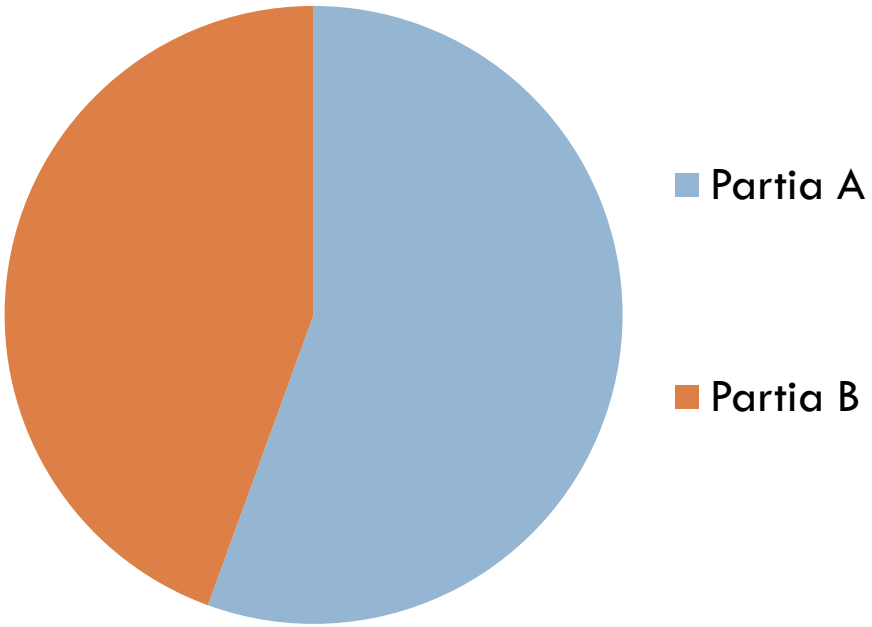
Problem braków danych

Logika badań reprezentacyjnych

Próba reprezentatywna

Populacja

n = 1000

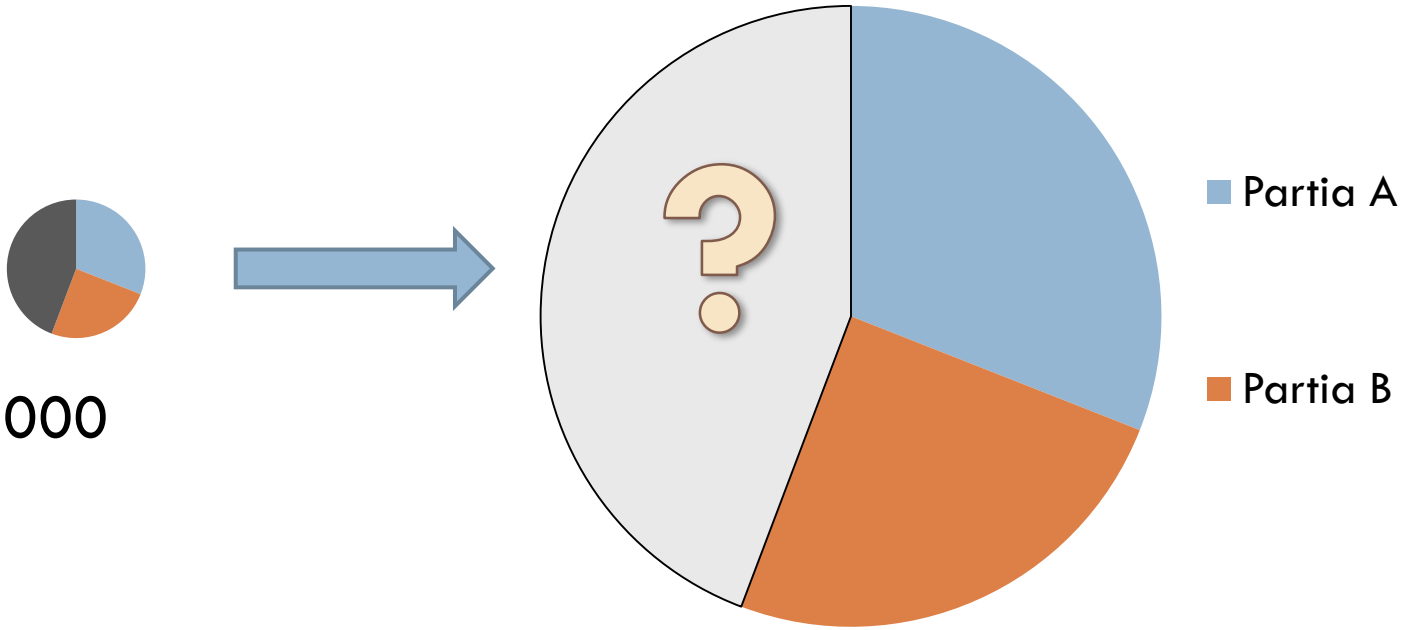


Gdy odsetek braków danych wynosi 0% mamy do czynienia z klasyczną sytuacją wnioskowania statystycznego.

Połowa długości przedziału ufności równa jest błędowi statystycznemu.

Problem braków danych

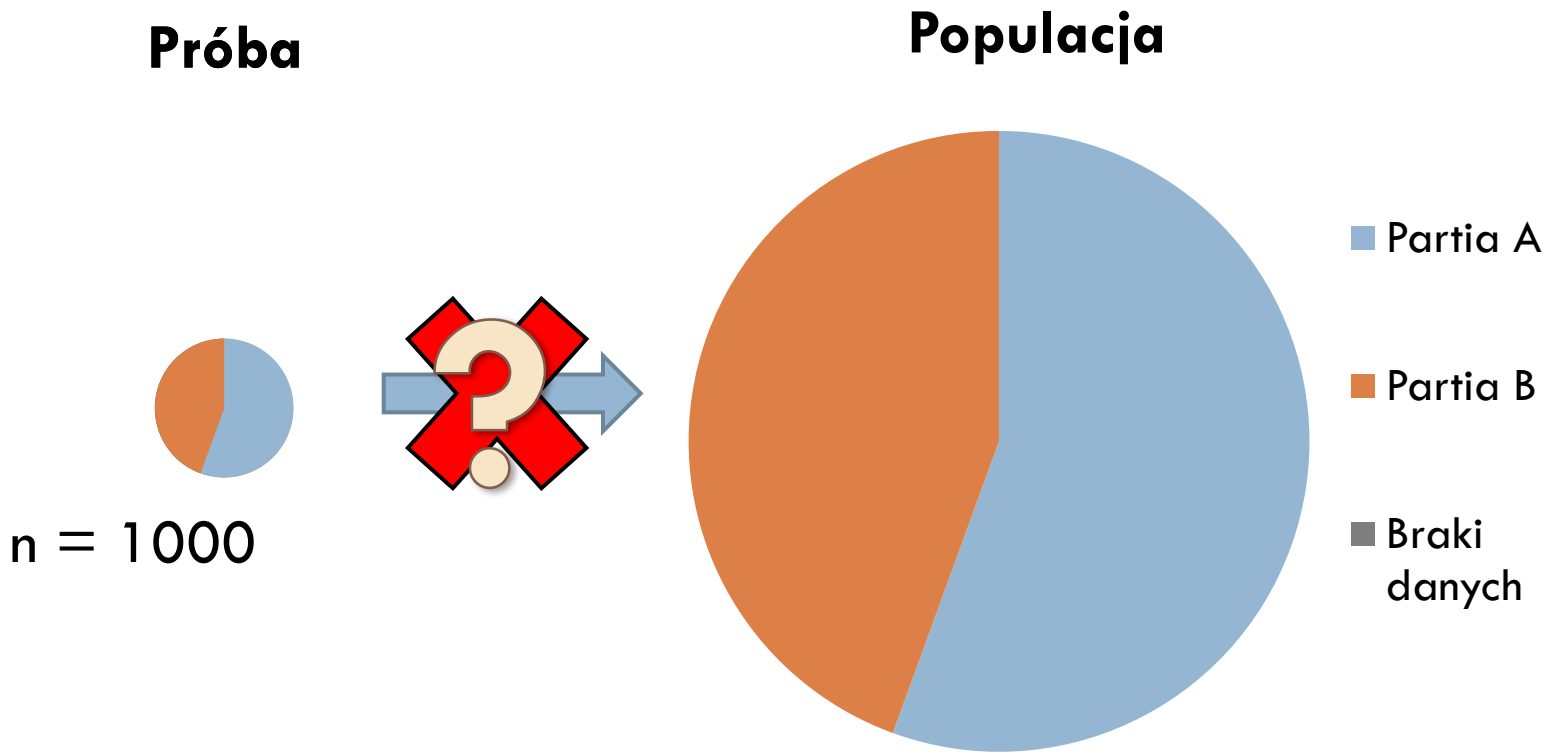
Próba **Populacja**



n = 1000

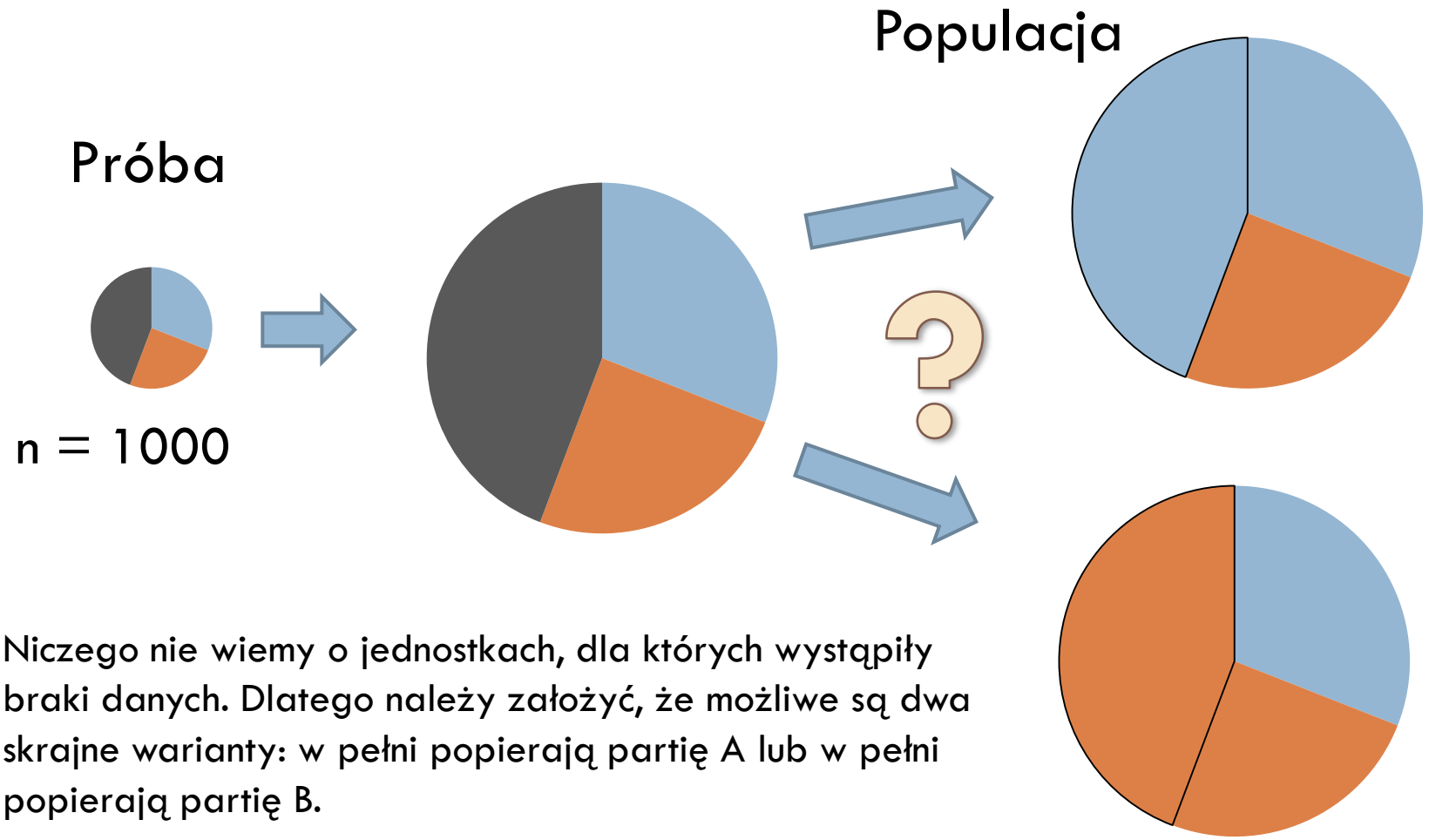
Gdy realizacja próby nie jest pełna powstaje pytanie: w jaki sposób należy wnioskować o części populacji odpowiadającej tej części próby, w której odnotowano braki danych?

Podejście „naiwne”



Nie uwzględnianie faktu, że części próby nie udało się zrealizować oznacza przyjęcie założenia, że jednostki niezbadane nie różnią się od jednostek zbadanych.
Jest to jednak założenie przynajmniej wątpliwe.

Podejście ostrożne



Niczego nie wiemy o jednostkach, dla których wystąpiły braki danych. Dlatego należy założyć, że możliwe są dwa skrajne warianty: w pełni popierają partię A lub w pełni popierają partię B.

Problem braków danych

- Zjawisko ma szeroki zasięg (zwłaszcza w badaniach CATI)
- Nie wiemy w jakiej mierzy błąd wynikający z problemu braków danych jest losowy, a w jakiej systematyczny.
- Dane i ostrożność każą przyjąć że jest on systematyczny.

Pseudo-przedział ufności

Na podstawie:

profesor Grzegorz Lissowski

Problem jednostek niedostępnych w reprezentacyjnych badaniach socjologicznych 1971

Szerokość przedziału ufności w przypadku występowania braków danych

Przedział ufności dla osób które odpowiedziały + Przedział ufności dla osób które nie odpowiedziały

G. Lissowski; *Problem jednostek niedostępnych w reprezentacyjnych badaniach socjologicznych;* Warszawa 1971.

Dwie warstwy

- Warstwa jednostek dostępnych
- Warstwa jednostek niedostępnych

- Jest to pewne uproszczenie, model.

  W rzeczywistości to czy dana jednostka ostatecznie udzieli odpowiedzi na pytanie zależy od wielu czynników i może mieć charakter probabilistyczny, a nie deterministyczny.

Środek pseudo-przedziału ufności

- Oszacowaniem średniej w populacji:

$$\text{Oszac. } \mu = w_1 E(x) + w_2 \frac{\mu_{2\min} - \mu_{2\max}}{2}$$

Oszac. μ – oszacowanie średniej populacyjnej,
$\mu_{2\,min}$, $\mu_{2\,max}$ – skrajne wartości jakie może osiągnąć średnia w populacji,
w_1 – frakcja jednostek dostępnych (w próbie),
w_2 – frakcja jednostek dostępnych (w próbie).

- W tym wystąpieniu pomijam dowód i od razu prezentuję wzory pozwalające wyliczyć oszacowania populacyjnych wartości na podstawie próby.

Szacowanie długości pseudo-przedziału ufności (dla doboru prostego niezależnego)

- Długość pseudo-przedziału ufności równa jest sumie długości przedziału ufności liczonego dla warstwy dostępnej i dla warstwy niedostępnej.

$$d = w_1 d_1 + w_2 d_2$$

d – połowa długości pseudo-przedziału ufności,
d_1 – połowa długości przedziału ufności wynikająca z błędu estymatora,
d_2 – połowa długości przedziału ufności wynikająca z błędu jednostek niedostępnych,
w_1 – frakcja jednostek dostępnych (w próbie),
w_2 – frakcja jednostek dostępnych (w próbie).

G. Lissowski; *Problem jednostek niedostępnych w reprezentacyjnych badaniach socjologicznych;* Warszawa 1971.

Szacowanie długości pseudo-przedziału ufności

- Dla jednostek dostępnych połowa jego długość wynosi:

$$d_1 = t_\alpha \frac{\sigma_1}{\sqrt{(n+1)w_1}}$$

$$d = t_\alpha \frac{\sigma}{\sqrt{n}}$$

n – zaplanowana liczebność próby,
σ_1 – odchylenie standardowe w zrealizowanej próbie,
t_α – funkcja odwrotna dystrybuanty dla prawdopodobieństwa $\frac{1-\alpha}{2}$, gdzie jest α założonym poziomem wiarygodności estymacji.

G. Lissowski; *Problem jednostek niedostępnych w reprezentacyjnych badaniach socjologicznych;* Warszawa 1971.

Szacowanie długości pseudo-przedziału ufności

- Dla ogólnego przypadku połowa długości przedziału ufności dla warstwy niedostępnej wynosi:

$$d_2 = \frac{\mu_{\min} + \mu_{\max}}{2}$$

μ_{min} – minimum jakie może osiągnąć średnia.
μ_{max} – maksimum jakie może osiągnąć średnia.

- Gdy nie posiadamy żadnej wiedzy o zakresie w jakim mieści się średnia w warstwie nie dostępnej, dla frakcji jego długość wynosi:

$$d_2 = \frac{0+1}{2} = \frac{1}{2}$$

G. Lissowski; *Problem jednostek niedostępnych w reprezentacyjnych badaniach socjologicznych;* Warszawa 1971.

Szacowanie długości pseudo-przedziału ufności

- W sumie dla całej populacji składającej się z warstw jednostek dostępnych i niedostępnych, połowa długości przedziału ufności równa jest sumie obu wielkości przeważonej przez udział warstw w populacji:

$$d = w_1 t_\alpha \frac{\sigma_1}{\sqrt{(n+1)w_1}} + w_2 \frac{\mu_{2\min} + \mu_{2\max}}{2}$$

G. Lissowski; *Problem jednostek niedostępnych w reprezentacyjnych badaniach socjologicznych;* Warszawa 1971.

Połowa szerokości przedziału ufności frakcji w przypadku występowania braków danych

	n_1=250	n_1=1000	n_1=2000	n_1=4000
w_1=100%	6,2%	3,1%	2,2%	1,5%
w_1=90%	11,2%	8,1%	7,2%	6,5%
w_1=80%	16,2%	13,1%	12,2%	11,5%
w_1=70%	21,2%	18,1%	17,2%	16,5%
w_1=60%	26,2%	23,1%	22,2%	21,5%
w_1=50%	31,2%	28,1%	27,2%	26,5%

w_1- udział warstwy dostępnej w populacji; 1 - odsetek braków danych.

Zakres wiedzy jaki zdobywa badacz w wyniku badania jest bezpośrednio związany z długością przedziału ufność. Dlatego często lepiej jest obniżyć n ale podnieść RR.

Czy to nie jest przesada?

- Klasyczny przedział ufności w sytuacji nie występowania braków danych mówi o zakresie w jakim z 95% prawdopodobieństwem mieści się szacowany parametr.
- Jeżeli chcemy zrealizować ten sam postulat (95% pewności) uwzględniając braki danych konieczne jest posługiwanie się pseudo-przedziałem ufności.
- Jedynym rozwiązaniem jest próba uściślenia zakresu w jakim mieści się średnia w warstwie niedostępnej.

Zakres średniej zmiennej wśród jednostek niedostępnych

- Jest on kluczowy dla rozmiaru błędu wynikającego z obecności braków danych.

$$d_2 = \frac{\mu_{\min} + \mu_{\max}}{2}$$

- Jeżeli udałoby się ograniczyć zakres możliwych wartości μ, to rozmiar całego błędu zostałby znacząco zredukowany.
- Jednak działanie takie wymaga posiadania informacji, które są bardzo trudne do uzyskania.

Przykład praktyczny

Przewidywanie wyniku referendum na podstawie PGSS

Problem: jaki byłby wynik referendum?

- Czy uwzględniając braki danych można przewidzieć wynik referendum?
- W kolejnych latach (1992, 1999, 2008) rośnie odsetek braków udziału.
- W każdym roku kilka procent (4-6%) odmawia udzielenia odpowiedzi.

PGSS – pytanie o poparcie dla prawa do aborcji

P102. Czy uważa Pan(i), że kobieta powinna mieć, czy też nie, możliwość legalnego przerwania ciąży: **[Ankieter: odczytać kolejny punkt pytania]**

Tak 1
Nie 2
NIE WIEM 8

a) jeśli istnieje wysokie prawdopodobieństwo, że dziecko urodzi się z poważnymi wadami? ☐

b) jeśli kobieta jest mężatką i nie chce mieć już więcej dzieci? ☐

c) jeśli ciąża poważnie zagraża zdrowiu kobiety? ☐

d) jeśli rodzina ma bardzo niskie dochody i nie może sobie pozwolić na więcej dzieci? ☐

e) jeśli ciąża była wynikiem gwałtu? ☐

f) jeśli kobieta nie jest mężatką i nie chce wyjść za mężczyznę z którym zaszła w ciążę? ☐

g) jeśli kobieta tak chce niezależnie od tego, jakie ma ku temu powody? ☐

PGSS – pytanie o poparcie dla prawa do aborcji

- Warianty oszacowań połowy długości przedziału ufności (d):

  I – klasyczny przedział (bez brania pod uwagę braków danych).

  II – pseudo-przedział ufności z wliczeniem braków udziału do braków danych.

  III – pseudo-przedział ufności z wliczeniem braków udziału oraz „nie wiem” do braków danych.

1992							
		ods. b.d.	d	E(x)	początek	koniec	konkluzywny
Wersja 1	Wariant I	0,0%	1,6%	88,4%	86,8%	90,0%	TAK
	Wariant II	17,7%	10,1%	81,6%	71,5%	91,7%	TAK
	Wariant III	23,6%	13,0%	79,4%	66,3%	92,4%	TAK
Wersja 2	Wariant I	0,0%	2,6%	53,3%	50,7%	55,9%	TAK
	Wariant II	17,7%	10,8%	52,7%	41,9%	63,5%	NIE
	Wariant III	29,3%	16,5%	52,3%	35,9%	68,8%	NIE
Wersja 3	Wariant I	0,0%	1,4%	90,9%	89,5%	92,4%	TAK
	Wariant II	17,7%	10,0%	83,7%	73,7%	93,7%	TAK
	Wariant III	23,9%	13,0%	81,2%	68,1%	94,2%	TAK
Wersja 4	Wariant I	0,0%	2,5%	61,4%	58,9%	64,0%	TAK
	Wariant II	17,7%	10,8%	59,4%	48,7%	70,2%	NIE
	Wariant III	28,6%	16,1%	58,2%	42,1%	74,3%	NIE
Wersja 5	Wariant I	0,0%	1,8%	85,5%	83,7%	87,3%	TAK
	Wariant II	17,7%	10,2%	79,2%	69,0%	89,4%	TAK
	Wariant III	26,3%	14,5%	76,2%	61,7%	90,6%	TAK
Wersja 6	Wariant I	0,0%	2,6%	46,8%	44,2%	49,4%	TAK
	Wariant II	17,7%	10,8%	47,4%	36,6%	58,2%	NIE
	Wariant III	30,1%	16,9%	47,8%	30,9%	64,6%	NIE
Wersja 7	Wariant I	0,0%	2,6%	44,6%	42,0%	47,2%	TAK
	Wariant II	17,7%	10,8%	45,6%	34,8%	56,4%	NIE
	Wariant III	29,1%	16,4%	46,2%	29,8%	62,5%	NIE

1999							
		ods. b.d.	d	E(x)	początek	koniec	konkluzywny
Wersja 1	Wariant I	0,0%	2,4%	82,0%	79,6%	84,4%	TAK
	Wariant II	33,4%	18,2%	71,3%	53,2%	89,5%	TAK
	Wariant III	40,6%	21,7%	69,0%	47,3%	90,7%	NIE
Wersja 2	Wariant I	0,0%	3,1%	43,0%	39,9%	46,1%	TAK
	Wariant II	33,4%	18,6%	45,3%	26,7%	63,9%	NIE
	Wariant III	43,0%	23,3%	46,0%	22,7%	69,3%	NIE
Wersja 3	Wariant I	0,0%	2,1%	86,7%	84,6%	88,8%	TAK
	Wariant II	33,4%	18,0%	74,4%	56,4%	92,4%	TAK
	Wariant III	41,0%	21,7%	71,6%	49,9%	93,4%	NIE
Wersja 4	Wariant I	0,0%	3,2%	49,7%	46,6%	52,9%	NIE
	Wariant II	33,4%	18,6%	49,8%	31,2%	68,4%	NIE
	Wariant III	43,9%	23,7%	49,9%	26,1%	73,6%	NIE
Wersja 5	Wariant I	0,0%	2,6%	77,6%	75,0%	80,2%	TAK
	Wariant II	33,4%	18,3%	68,4%	50,1%	86,7%	TAK
	Wariant III	42,5%	22,8%	65,9%	43,1%	88,6%	NIE
Wersja 6	Wariant I	0,0%	3,0%	34,1%	31,1%	37,1%	TAK
	Wariant II	33,4%	18,5%	39,4%	20,9%	57,9%	NIE
	Wariant III	43,9%	23,6%	41,1%	17,5%	64,7%	NIE
Wersja 7	Wariant I	0,0%	3,0%	32,6%	29,6%	35,5%	TAK
	Wariant II	33,4%	18,5%	38,4%	19,9%	56,9%	NIE
	Wariant III	43,7%	23,5%	40,2%	16,7%	63,7%	NIE

2008							
		ods. b.d.	d	E(x)	początek	koniec	konkluzywny
Wersja 1	Wariant I	0,0%	2,1%	83,8%	81,7%	85,9%	TAK
	Wariant II	48,2%	25,1%	67,5%	42,4%	92,6%	NIE
	Wariant III	52,3%	27,1%	66,1%	39,0%	93,3%	NIE
Wersja 2	Wariant I	0,0%	2,8%	36,6%	33,8%	39,3%	TAK
	Wariant II	48,2%	25,4%	43,0%	17,6%	68,5%	NIE
	Wariant III	53,3%	27,9%	43,7%	15,8%	71,7%	NIE
Wersja 3	Wariant I	0,0%	1,7%	89,6%	87,8%	91,3%	TAK
	Wariant II	48,2%	25,0%	70,5%	45,6%	95,5%	NIE
	Wariant III	52,0%	26,8%	69,0%	42,2%	95,8%	NIE
Wersja 4	Wariant I	0,0%	2,8%	39,5%	36,7%	42,4%	TAK
	Wariant II	48,2%	25,5%	44,6%	19,1%	70,0%	NIE
	Wariant III	53,7%	28,1%	45,2%	17,0%	73,3%	NIE
Wersja 5	Wariant I	0,0%	2,3%	81,0%	78,7%	83,3%	TAK
	Wariant II	48,2%	25,2%	66,1%	40,9%	91,3%	NIE
	Wariant III	53,6%	27,8%	64,4%	36,5%	92,2%	NIE
Wersja 6	Wariant I	0,0%	2,6%	27,5%	24,9%	30,0%	TAK
	Wariant II	48,2%	25,3%	38,3%	13,0%	63,7%	NIE
	Wariant III	53,3%	27,9%	39,5%	11,6%	67,3%	NIE
Wersja 7	Wariant I	0,0%	2,5%	24,2%	21,8%	26,7%	TAK
	Wariant II	48,2%	25,3%	36,7%	11,4%	62,0%	NIE
	Wariant III	53,4%	27,8%	38,0%	10,2%	65,8%	NIE

Konkluzje

- Wykorzystywanie metody pseudo-przedziału ufności skłania do wnioski, że braki danych to ogromny problem dzisiejszych badań społecznych.
- Wnioski z wielu badań społecznych można poddać w wątpliwość zwracając uwagę na braki danych.
- Konieczna jest praca nad:
  - metodami wolnymi od braków danych,
  - zbadaniem w jakiej mierze braki danych powodują błędy losowe, a w jakiej systematyczne.

DZIĘKUJĘ ZA UWAGĘ
Z CHĘCIĄ UDZIELĘ ODPOWIEDZI NA PYTANIA

2013-12-15

Piotr Zielonka