

# Estymacja na podstawie prób nielosowych

dr Maciej Beręsewicz, prof. UEP

Katedra Statystyki, Uniwersytet Ekonomiczny w Poznaniu  
Ośrodek Metodologii Badań Ludnościowych, Urząd Statystyczny  
w Poznaniu

# Spis treści

- 1 O wykładzie
- 2 Internet w Polsce
- 3 Google Trends
- 4 Web scraping
- 5 Reprezentatywność
  - Reprezentatywność – definicje, pomiar, problematyka
- 6 Metody estymacji dla prób nielosowych
  - Podstawowe założenia na potrzeby zajęć
  - Metody quasi-randomizacyjne
  - Metody oparte na modelu
  - Podwójnie odporne estymatory

# Spis treści

- 1 O wykładowie
- 2 Internet w Polsce
- 3 Google Trends
- 4 Web scraping
- 5 Reprezentatywność
- 6 Metody estymacji dla prób nielosowych

# Literatura (wybrana)

- Baker, R, J Michael Brick, Nancy A Bates, Mike Battaglia, Mick P Couper, Jill A Dever, Krista J Gile, and Roger Tourangeau (2013). Summary Report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology* 1, pp. 90–143.
- Bethlehem, J. (2010). Selection Bias in Web Surveys. *International Statistical Review*, 78(2), 161–188. doi:10.1111/j.1751-5823.2010.00112.x.
- Bethlehem, J., and Biggignandi, S. (2012). *Handbook of Web Surveys*, John Wiley and Sons, Inc. doi:10.1086/318641.
- Callegaro M., Baker R., Bethlehem J., Göritz A.S., Krosnick J.A., Lavrakas P. J. (2014) *Online Panel Research A Data Quality Perspective*, Wiley.
- Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, 22(2), 329–349.
- Kim, J. K., and Tam, S. M. (2020). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review*.
- S. Yang, J.K. Kim, and R. Song (2020). "Doubly Robust Inference when Combining Probability and Non-probability Samples with High-dimensional Data", *Journal of the Royal Statistical Society: Series B*, 82, 445-465.
- J.K. Kim and Z. Wang (2019). "Sampling techniques for big data analysis in finite population inference", *International Statistical Review*, 87, S177-S191.

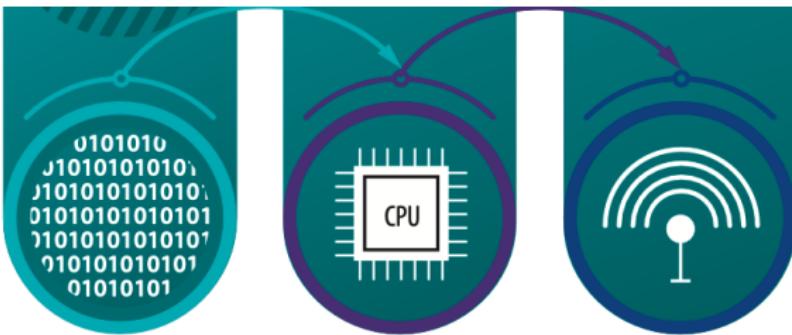
# Spis treści

- 1 O wykładzie
- 2 Internet w Polsce
- 3 Google Trends
- 4 Web scraping
- 5 Reprezentatywność
- 6 Metody estymacji dla prób nielosowych

# Internet w Polsce – dane GUS i Eurostat

- Households – level of internet access
- Households – type of connection to the internet
- Households with access to the internet at home
- Individuals – frequency of internet use
- Individuals – internet activities
- Individuals – internet use
- Individuals using the internet for interacting with public authorities
- Individuals who ordered goods or services over the internet for private use
- Individuals who used the internet for interaction with public authorities
- Individuals who used the internet, frequency of use and activities
- Type of connections to the internet
- Use of computers and the internet by employees
- Use of mobile connections to the internet
- Use of mobile connections to the internet by employees

# Internet w Polsce – źródło: badanie ICT GUS



## Społeczeństwo informacyjne w Polsce w 2024 r.

Information society in Poland in 2024

# Internet w Polsce – źródło: badanie ICT GUS

## Gospodarstwa domowe posiadające dostęp do Internetu w domu

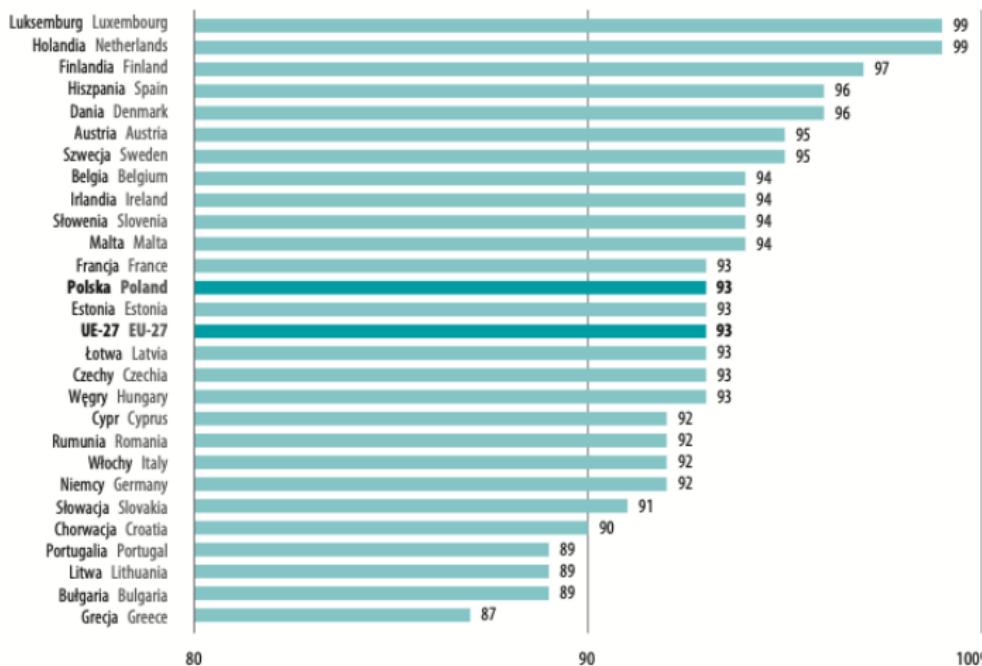
Households with access to the Internet at home

Wyszczególnienie Specification	2020	2021	2022	2023	2024
	w % ogółu gospodarstw danej grupy in % of total households in a group				
<b>Ogółem Total</b>	<b>90,4</b>	<b>92,4</b>	<b>93,3</b>	<b>93,3</b>	<b>95,9</b>
Typ gospodarstwa domowego Household type					
Gospodarstwa z dziećmi Households with children	99,5	99,7	99,9	99,8	99,9
Gospodarstwa bez dzieci Households without children	85,9	88,8	90,5	90,3	94,3
Miejsce zamieszkania Domicile					
Duże miasta Large cities	92,1	93,8	94,4	94,8	96,7
Mniejsze miasta Small cities	89,7	91,6	92,3	92,0	95,6
Obszary wiejskie Rural areas	89,3	91,8	93,2	93,2	95,5
Stopień urbanizacji Degree of urbanisation					
Niski Thinly populated	88,9	91,9	92,8	92,9	95,3
Średni Intermediate density	90,4	91,1	92,6	92,5	95,5
Wysoki Densely populated	91,6	93,7	94,2	94,2	96,6
Obszary Areas					
Polska wschodnia Eastern Poland	88,9	90,0	91,7	92,1	95,2
Polska centralna Central Poland	90,8	93,1	94,2	94,1	96,1
Polska zachodnia Western Poland	90,6	93,1	92,7	92,6	96,0

# Internet w Polsce – źródło: badanie ICT GUS

**Gospodarstwa domowe z dostępem do Internetu w domu w krajach Unii Europejskiej w 2023 r.**

Households with access to the Internet at home in European Union countries in 2023



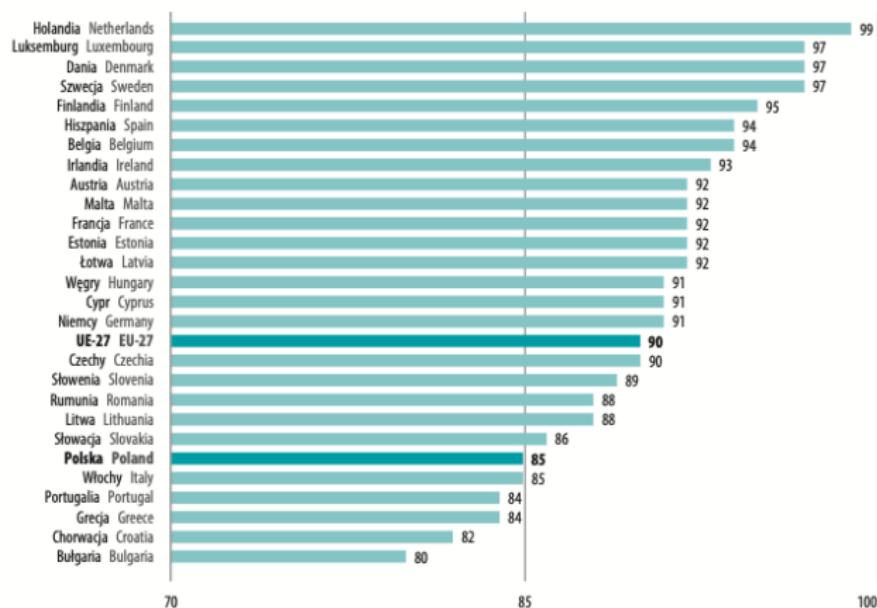
# Internet w Polsce – źródło: badanie ICT GUS

## Gospodarstwa domowe posiadające szerokopasmowy dostęp do Internetu w domu Households with broadband access to the Internet at home

Wyszczególnienie Specification	2020	2021	2022	2023	2024
	w % ogółu gospodarstw danej grupy in % of total households in a group				
<b>Ogółem Total</b>	<b>89,6</b>	<b>91,7</b>	<b>92,6</b>	<b>92,8</b>	<b>94,7</b>
Typ gospodarstwa domowego Household type					
Gospodarstwa z dziećmi Households with children	99,1	99,4	99,6	99,7	99,5
Gospodarstwa bez dzieci Households without children	84,9	87,9	89,6	89,6	92,8
Miejsce zamieszkania Domicile					
Duże miasta Large cities	91,0	93,1	93,4	93,9	95,2
Mniejsze miasta Small cities	89,1	91,2	91,6	91,7	94,9
Obszary wiejskie Rural areas	88,7	90,9	92,8	92,8	94,1
Stopień urbanizacji Degree of urbanisation					
Niski Thinly populated	88,3	91,1	92,5	92,5	94,3
Średni Intermediate density	89,7	90,6	91,8	92,2	94,3
Wysoki Densely populated	90,5	93,0	93,3	93,4	95,2
Obszary Areas					
Polska wschodnia Eastern Poland	88,4	89,6	91,3	91,6	95,0
Polska centralna Central Poland	90,1	92,3	93,6	93,6	94,8
Polska zachodnia Western Poland	89,2	92,4	91,5	92,0	94,2

# Internet w Polsce – źródło: badanie ICT GUS

**Osoby regularnie korzystające z Internetu w krajach Unii Europejskiej w 2023 r.**  
Regular Internet users in European Union countries in 2023



# Internet w Polsce – źródło: badanie ICT GUS

## Częstotliwość korzystania z Internetu

Frequency of Internet use

Wyszczególnienie Specification	2020	2021	2022	2023	2024
-----------------------------------	------	------	------	------	------

W % ogółu osób In % of total individuals

Regularnie Regularly	81,4	83,6	85,7	85,3	87,6
Codziennie lub prawie codziennie Every day or almost every day	72,3	73,7	80,3	79,5	83,9
Kilka razy dziennie Several times during the day	66,7	67,6	64,4	62,2	70,7
Przynajmniej raz w tygodniu, ale nie każdego dnia At least once a week but not every day	9,0	10,0	5,4	5,8	3,7

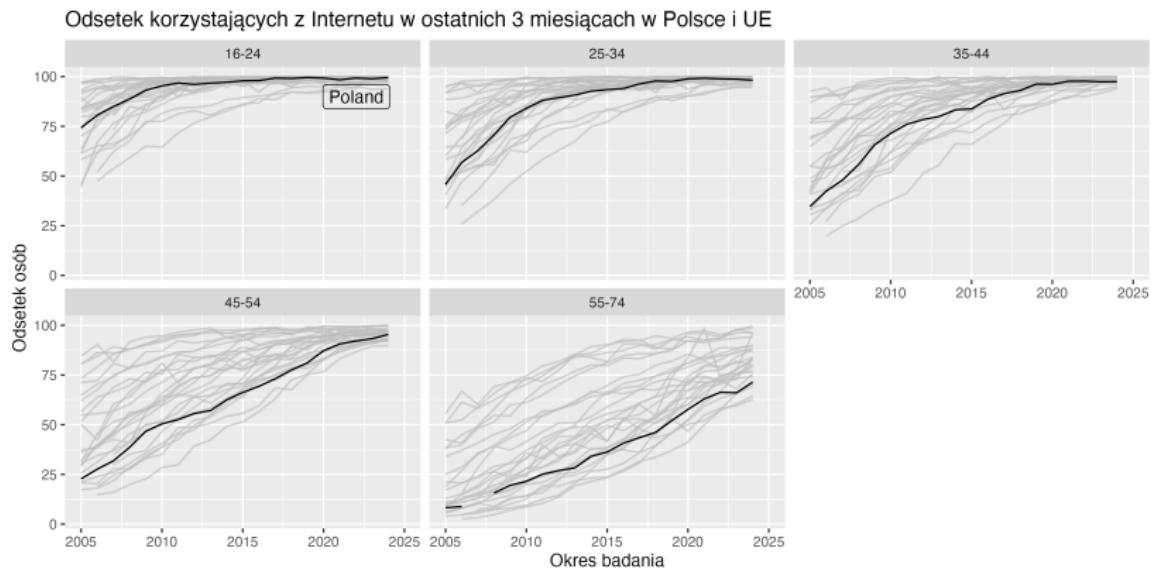
## Wyszczególnienie Specification

2019 2020 2021 2022 2023

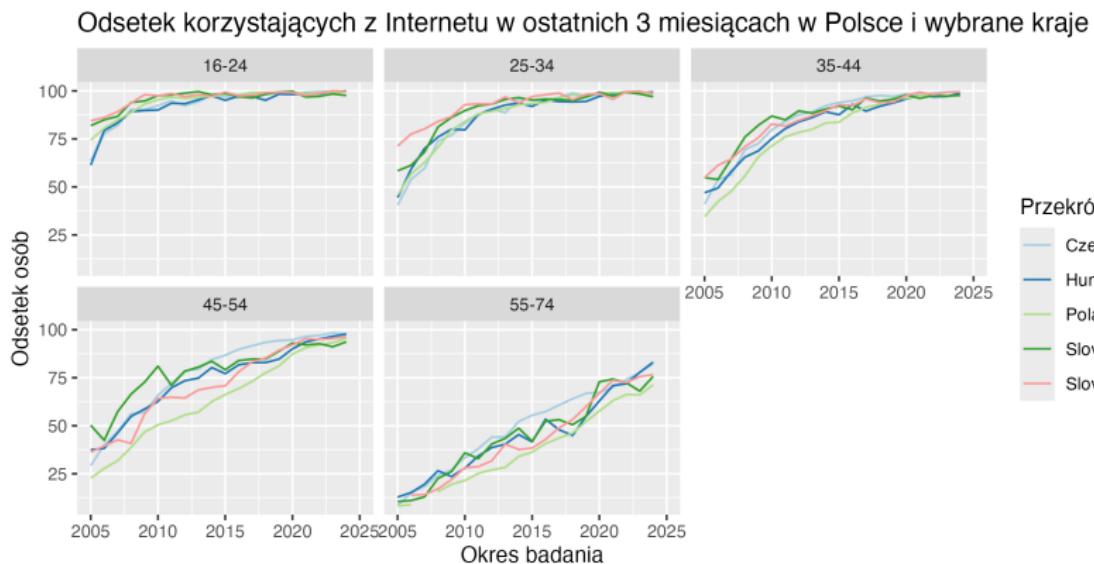
W % osób korzystających z Internetu w ciągu ostatnich 3 miesięcy  
In % of individuals using the internet in the last 3 months

Regularnie Regularly	97,3	97,8	98,0	98,6	98,7
Codziennie lub prawie codziennie Every day or almost every day	84,8	87,0	86,3	92,4	92,0
Przynajmniej raz w tygodniu, ale nie każdego dnia At least once a week but not every day	12,6	10,9	11,7	6,2	6,7
Rzadziej niż raz w tygodniu Less than once a week	2,7	2,2	2,0	1,4	1,3

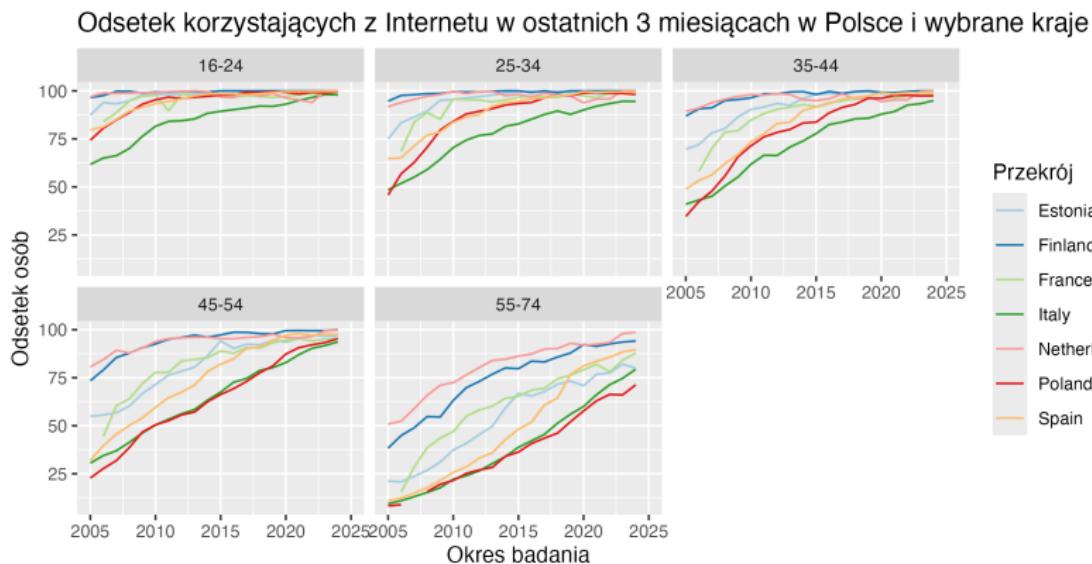
# Internet w Polsce – źródło: badanie ICT GUS



# Internet w Polsce – źródło: badanie ICT GUS



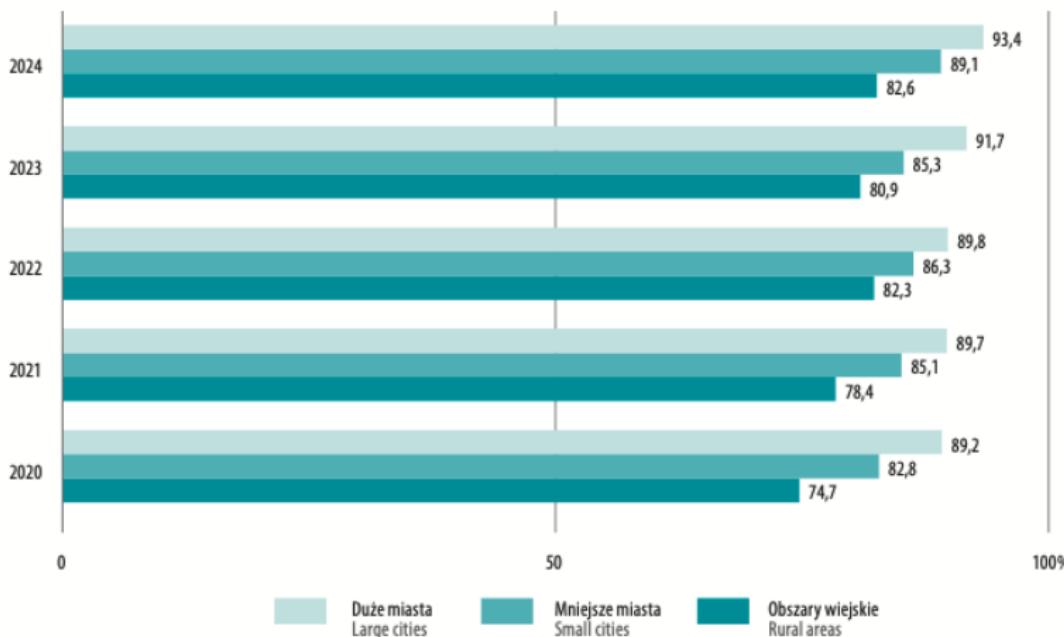
# Internet w Polsce – źródło: badanie ICT GUS



# Internet w Polsce – źródło: badanie ICT GUS

## Osoby regularnie korzystające z Internetu według miejsca zamieszkania

Regular Internet users by domicile



# Internet w Polsce – źródło: badanie ICT GUS

## Osoby regularnie korzystające z Internetu według grup wieku

Regular Internet users by age groups

Wyszczególnienie Specification		2020	2021	2022	2023	2024
		w % ogółu osób danej grupy in % of total individuals in a group				
16–24 lata	16–24 years	99,2	98,4	99,0	98,9	99,5
25–34		98,4	98,9	98,7	98,6	98,1
35–44		95,2	96,7	97,1	97,1	97,3
45–54		84,3	89,1	91,1	92,3	94,4
55–64		65,8	71,3	75,5	74,6	82,5
65–74 lata	65–74 years	40,4	45,9	51,0	51,9	56,8

# Internet w Polsce – źródło: badanie ICT GUS

**Gospodarstwa domowe posiadające dostęp do Internetu w domu oraz osoby korzystające z Internetu według województw w 2024 r.**

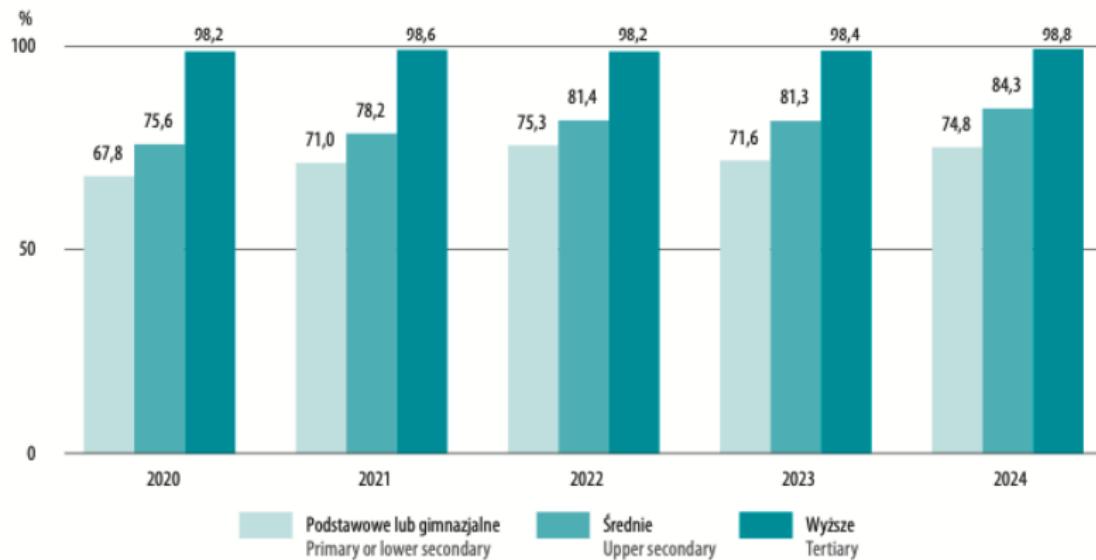
Households with access to the Internet at home and Internet users by voivodships in 2024

Województwa Voivodships	Odsetek gospodarstw domowych posiadających dostęp do Internetu w domu Percentage of households with access to the Internet at home	Odsetek osób korzystających z Internetu Percentage of individuals using the Internet	Odsetek osób regularnie korzystających z Internetu Percentage of regular internet users
<b>Polska Poland</b>	<b>95,9</b>	<b>91,5</b>	<b>87,6</b>
Dolnośląskie	96,1	92,8	90,2
Kujawsko-pomorskie	97,5	89,3	87,0
Lubelskie	93,6	90,6	85,3
Lubuskie	96,4	91,8	87,2
Łódzkie	94,6	92,2	88,5
Małopolskie	95,7	92,0	85,9
Mazowieckie	96,5	93,4	90,5
Opolskie	95,1	85,4	79,5
Podkarpackie	98,5	90,9	84,5
Podlaskie	94,8	86,9	80,4
Pomorskie	95,4	91,4	89,6
Śląskie	96,3	94,0	92,1
Świętokrzyskie	94,6	85,9	80,5
Wielkopolskie	97,2	96,5	92,6

# Internet w Polsce – źródło: badanie ICT GUS

## Osoby regularnie korzystające z Internetu według poziomu wykształcenia

Regular Internet users by education level



# Internet w Polsce – źródło: badanie ICT GUS

## Osoby regularnie korzystające z Internetu według aktywności zawodowej

Regular Internet users by employment situation

Wyszczególnienie Specification	2020	2021	2022	2023	2024
	w % ogółu osób danej grupy in % of total individuals in a group				
Emeryci i inni bierni zawodowo Retired or other not in the labour force	53,5	57,6	60,4	62,4	66,6
Bezrobotni Unemployed	82,6	83,2	86,3	86,7	87,5
Pracujący Persons employed	91,8	93,3	95,2	94,5	96,5
Rolnicy Farmers	65,1	75,1	83,0	81,2	84,4
Pracujący na własny rachunek Self-employed	96,0	98,0	96,7	96,3	93,7
Pracownicy najemni Employees	93,6	94,8	96,1	95,5	97,2
Uczniowie i studenci Students	99,8	99,1	99,3	99,8	99,9

# Internet w Polsce – źródło: badanie ICT GUS

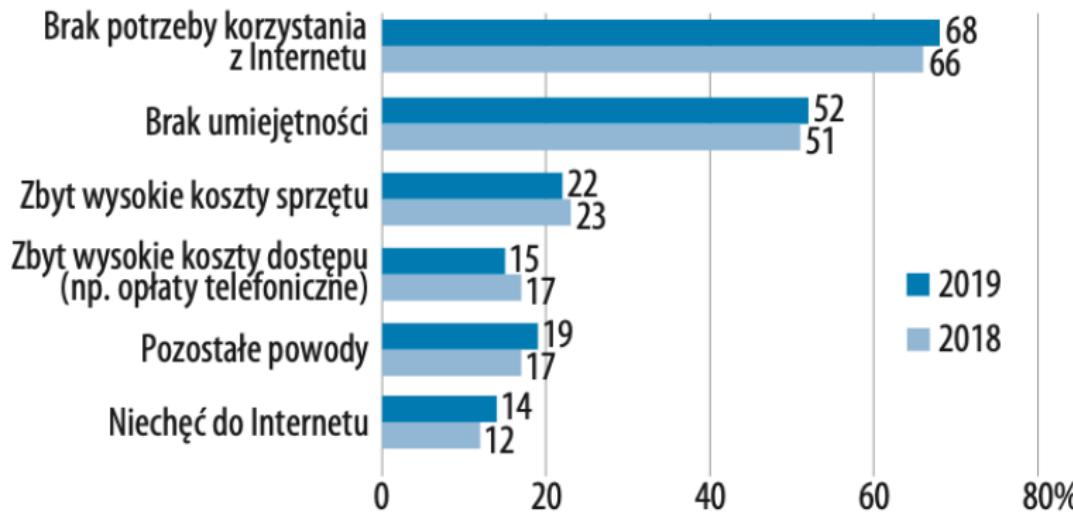
## Osoby korzystające z Internetu w sprawach prywatnych w ciągu ostatnich 3 miesięcy według wybranych celów

Individuals using the Internet for private purposes in the last 3 months by selected activities

Cele korzystania z Internetu Purposes of Internet usage	2018	2019	2020	2021	2022	2018	2019	2020	2021	2022
	w % ogółu osób in % of total individuals					w % osób korzystających z internetu in % of Internet users				
Korzystanie z poczty elektronicznej Sending, receiving e-mail	60,7	64,8	65,9	68,3	69,3	78,2	80,6	79,2	80,0	79,7
Wyszukiwanie informacji o towarach i usługach Finding information about goods and services	64,0	62,2	62,7	65,6	74,3	82,5	77,4	75,4	76,9	85,4
Czytanie online wiadomości, gazet lub czasopism Reading online news, newspapers or magazines	.	60,5	65,4	69,4	64,3	.	75,2	78,6	81,3	73,9
Korzystanie z serwisów społecznościowych Participating in social networks	49,9	53,0	54,8	56,8	60,6	64,3	65,9	65,9	66,5	69,7
Korzystanie z komunikatorów Using instant messaging	-	48,6	53,4	58,5	64,7	-	60,4	64,2	68,5	74,4
Korzystanie z usług bankowym Internet banking	44,0	47,3	49,5	52,2	55,6	56,8	58,8	59,5	61,2	63,9
Wykonywanie rozmów głosowych lub video przez Internet Making calls (including video calls) over the internet	34,1	48,6	55,0	56,4	54,8	44,0	60,4	66,1	66,1	63,0
Oglądanie nagrani video z serwisów tworzonych przez użytkowników (np. YouTube) Watching video content from sharing services	44,6	41,6	41,5	42,4	47,8	57,5	51,7	49,8	49,7	55,0

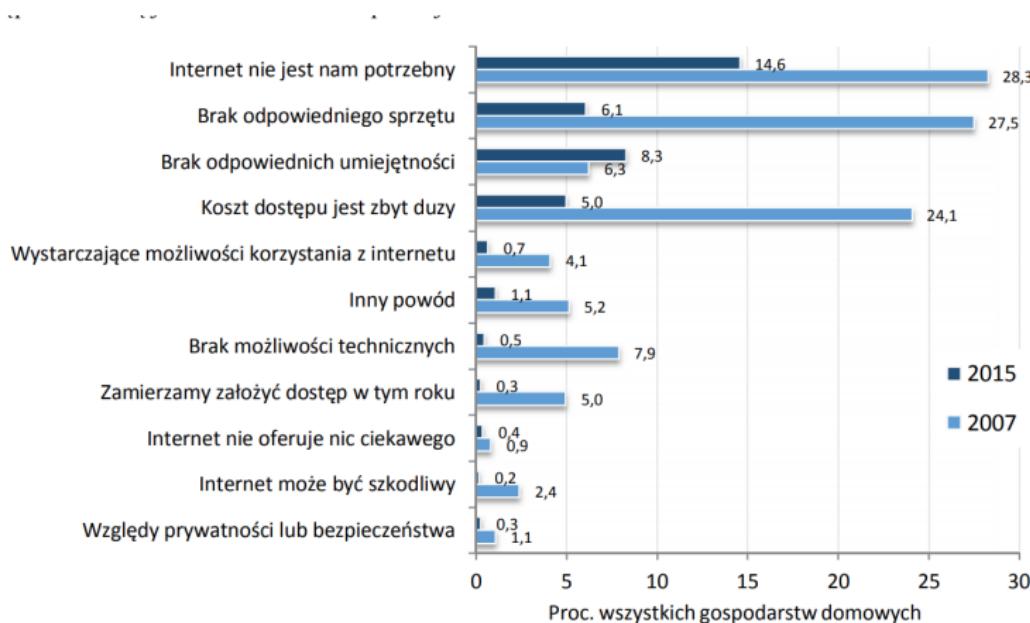
## Internet w Polsce – źródło: badanie ICT GUS

Z jakich powodów 13% gospodarstw domowych w Polsce nie miało dostępu do Internetu w 2019 r. (w porównaniu z sytuacją w 2018 r.)



W procentach ogółu gospodarstw domowych bez dostępu do Internetu.

# Internet w Polsce – źródło: Diagnoza Społeczna 2015

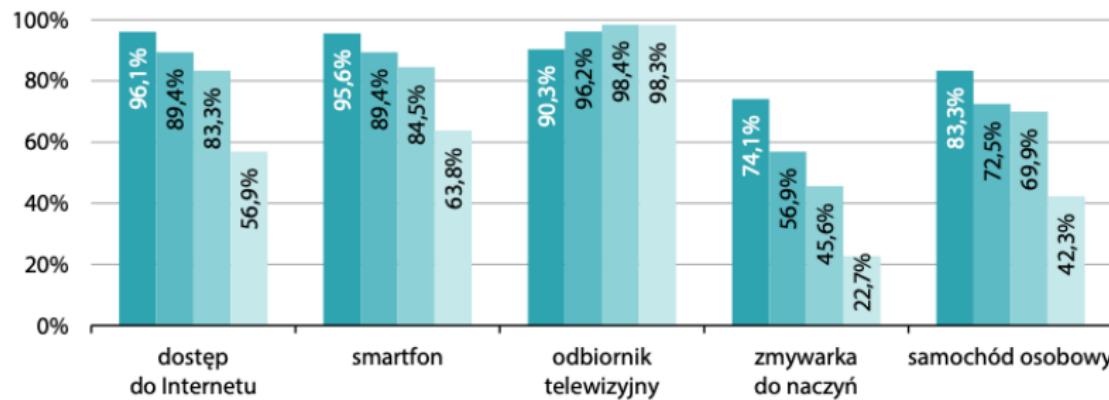


Wykres 7.1.7. Powody braku dostępu do internetu w gospodarstwach domowych w latach 2007 i 2015.

# Internet w Polsce – źródło: BBGD 2022

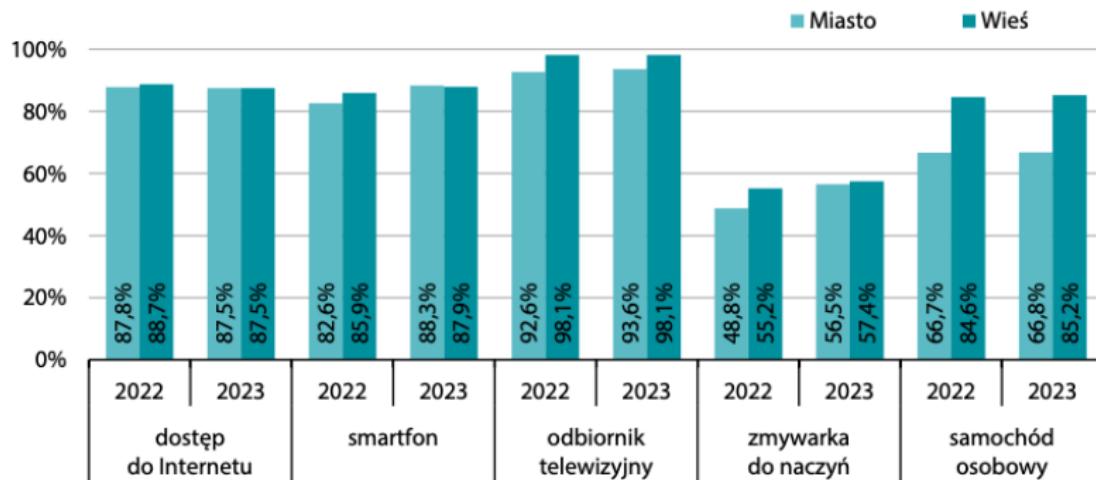
**Wypożyczenie gospodarstw domowych w wybrane dobra trwałego użytkowania według poziomu wykształcenia osoby odniesienia w 2023 r.**

- wyższe
- policealne, średnie (ogólnokształcące, zawodowe/branżowe)
- zasadnicze (zawodowe/branżowe)
- gimnazjalne, podstawowe, bez wykształcenia



# Internet w Polsce – źródło: BBGD 2022

**Wyposażenie gospodarstw domowych w wybrane dobra trwałego użytkowania według miejsca zamieszkania w latach 2022–2023**



# Internet w Polsce – literatura

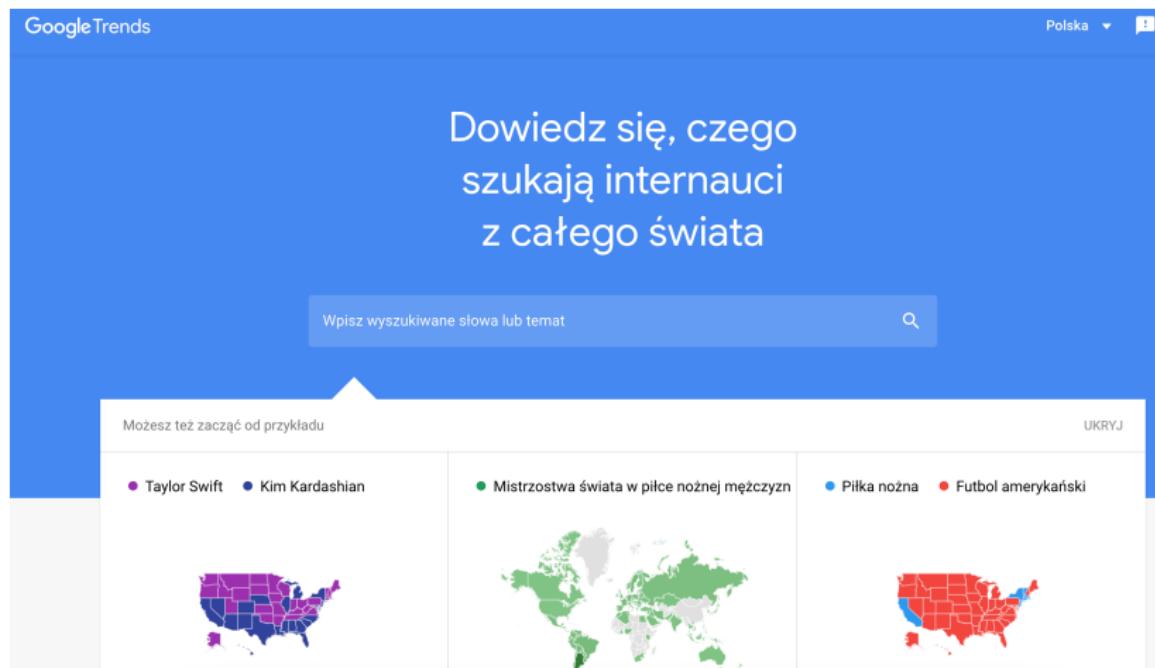
Literatura użyta na potrzeby zajęć o Internecie w Polsce

- GUS (2023), Społeczeństwo informacyjne w Polsce w 2022 r.,  
<https://stat.gov.pl/obszary-tematyczne/nauka-i-technika-spoleczenstwo-informacyjne/spoleczenstwo-informacyjne/spoleczenstwo-informacyjne-w-polsce-w-2022-roku,1,16.html>.
- GUS(2022), Jak korzystamy z Internetu? 2022,  
<https://stat.gov.pl/obszary-tematyczne/nauka-i-technika-spoleczenstwo-informacyjne/spoleczenstwo-informacyjne/jak-korzystamy-z-internetu-2022,5,13.html>
- GUS (2021), Budżety gospodarstw domowych w 2019 roku,  
<https://stat.gov.pl/obszary-tematyczne/warunki-zycia/dochody-wydatki-i-warunki-zycia-ludnosci/budzety-gospodarstw-domowych-w-2019-roku,9,14.html>
- Czapiewski (2016) Raport z badania Diagnoza Społeczna 2015,  
[www.diagnoza.com](http://www.diagnoza.com)

# Spis treści

- 1 O wykładowie
- 2 Internet w Polsce
- 3 Google Trends
- 4 Web scraping
- 5 Reprezentatywność
- 6 Metody estymacji dla prób nielosowych

# Google Trends

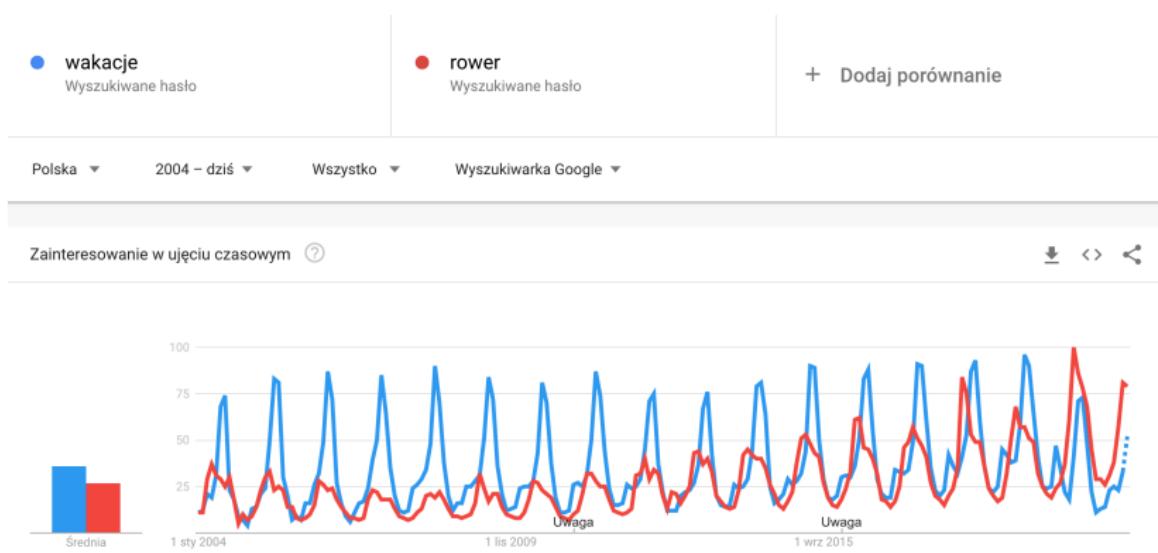


Rysunek 1: Strona główna Google Trends

# Google Trends – podstawowe informacje

- 11 maj 2006 – start jako "Google Insights for Search".
- 27 września 2012 – powstanie Google Trends w formie, którą teraz znamy.
- Trendy Google zapewniają dostęp do zasadniczo niefiltrowanej próbki rzeczywistych żądań wyszukiwania wysyłanych do Google.
- Dane te są zanonimizowane (nie można rozpoznać użytkowników), podzielone na kategorie (z określonym tematem wyszukiwanych haseł) i zagregowane (pogrupowane).
- W Trendach Google dostępne są dwa rodzaje próbek danych:
  - Dane w czasie rzeczywistym to próbka obejmująca ostatnie siedem dni.
  - Dane niedynamiczne stanowią osobną próbkę i obejmują okres od 2004 roku do 36 godzin przed wyszukiwaniem.
- Dane Google Trends podlegają normalizacji.

# Google Trends – normalizacja



Rysunek 2: Przykład zapytania i normalizacji

# Google Trends – normalizacja

Trendy Google normalizują dane wyszukiwania, by ułatwić porównywanie haseł. Wyniki wyszukiwania są normalizowane względem czasu i lokalizacji wyszukiwania w ten sposób:

- Każdy punkt danych jest dzielony przez łączną liczbę wyszukiwań w danym regionie i okresie, co pozwala ocenić jego względną popularność. W przeciwnym razie regiony, w których wyszukiwań jest najwięcej, zawsze byłyby najwyższe w rankingu.
- Rezultat jest następnie skalowany w zakresie od 0 do 100 na podstawie proporcjonalności tematu względem wszystkich wyszukiwań wszystkich tematów.
- Jeśli w różnych regionach zainteresowanie wyszukiwaniem wybranego hasła jest takie samo, nie oznacza to jeszcze, że łączna liczba wyszukiwań jest w nich również taka sama.

# Google Trends – jakość danych

- Dane w Trendach Google odzwierciedlają codzienne wyszukiwania w Google.
- Mogą też pokazywać nieprawidłową aktywność związaną z wyszukiwaniem, np. wyszukiwanie zautomatyzowane lub zapytania potencjalnie związane z próbami spamowania wyników wyszukiwania.
- Trendy Google odfiltrowują niektóre typy wyszukiwań, między innymi:
  - Wyszukiwania przeprowadzone przez niewielu użytkowników: Trendy pokazują tylko dane o hasłach, które są często wyszukiwane. Hasła o małej liczbie wyszukiwań są oznaczone jako „0”.
  - Powtarzające się wyszukiwania: Trendy pomijają identyczne zapytania wpisywane przez tego samego użytkownika w krótkich odstępach czasu.
  - Znaki specjalne: Trendy odfiltrowują zapytania z apostrofami i innymi znakami specjalnymi.

# Google Trends – przykłady



## Predicting the Present with Google Trends

HYUNYOUNG CHOI and HAL VARIAN

*Google, Inc., California, USA*

*In this paper we show how to use search engine data to forecast near-term values of economic indicators. Examples include automobile sales, unemployment claims, travel destination planning and consumer confidence.*

# Google Trends – przykłady

INVITED ARTICLE

SURFING THE WEB

Victor L. Yu, Section Editor

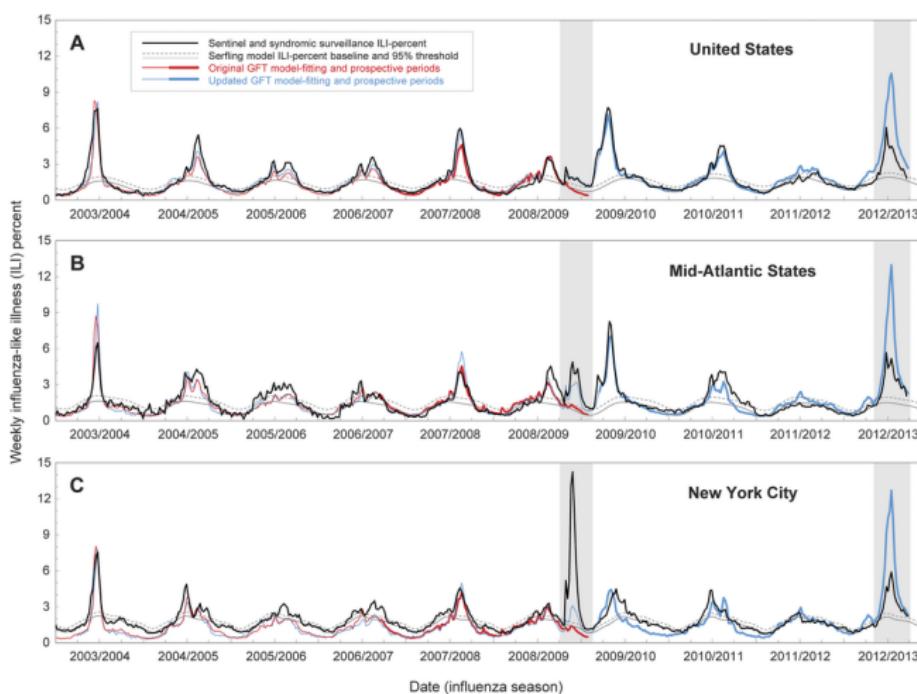
## Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks

**Herman Anthony Carneiro<sup>1,2</sup> and Eleftherios Mylonakis<sup>1</sup>**

<sup>1</sup>Division of Infectious Diseases, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts; and <sup>2</sup>School of Medicine, Imperial College London, London, United Kingdom

Google Flu Trends can detect regional outbreaks of influenza 7–10 days before conventional Centers for Disease Control and Prevention surveillance systems. We describe the Google Trends tool, explain how the data are processed, present examples, and discuss its strengths and limitations. Google Trends shows great promise as a timely, robust, and sensitive surveillance system. It is best used for surveillance of epidemics and diseases with high prevalences and is currently better suited to track disease activity in developed countries, because to be most effective, it requires large populations of Web search users. Spikes in search volume are currently hard to interpret but have the benefit of increasing vigilance. Google should work with public health care practitioners to develop specialized tools, using Google Flu Trends as a blueprint, to track infectious diseases. Suitable Web search query proxies for diseases need to be established for specialized tools or syndromic surveillance. This unique and innovative technology takes us one step closer to true real-time outbreak surveillance.

# Google Trends – przykłady



# Google Trends – przykłady



## Quantifying Trading Behavior in Financial Markets Using *Google Trends*

Tobias Preis<sup>1\*</sup>, Helen Susannah Moat<sup>2,3\*</sup> & H. Eugene Stanley<sup>2\*</sup>

<sup>1</sup>Warwick Business School, University of Warwick, Scarman Road, Coventry, CV4 7AL, UK, <sup>2</sup>Department of Physics, Boston University, 590 Commonwealth Avenue, Boston, Massachusetts 02215, USA, <sup>3</sup>Department of Civil, Environmental and Geomatic Engineering, UCL, Gower Street, London, WC1E 6BT, UK.

SUBJECT AREAS:  
STATISTICAL PHYSICS,  
THERMODYNAMICS AND  
NONLINEAR DYNAMICS  
APPLIED PHYSICS  
COMPUTATIONAL SCIENCE  
INFORMATION THEORY AND  
COMPUTATION

Received  
25 February 2013

Accepted  
3 April 2013

Crises in financial markets affect humans worldwide. Detailed market data on trading decisions reflect some of the complex human behavior that has led to these crises. We suggest that massive new data sources resulting from human interaction with the Internet may offer a new perspective on the behavior of market participants in periods of large market movements. By analyzing changes in *Google* query volumes for search terms related to finance, we find patterns that may be interpreted as “early warning signs” of stock market moves. Our results illustrate the potential that combining extensive behavioral data sets offers for a better understanding of collective human behavior.

# Google Trends – przykłady

*Journal of Forecasting*

*J. Forecast.* **30**, 565–578 (2011)

Published online 13 January 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/for.1213

## Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends

SIMEON VOSEN\* AND TORSTEN SCHMIDT

RWI, Essen, Germany

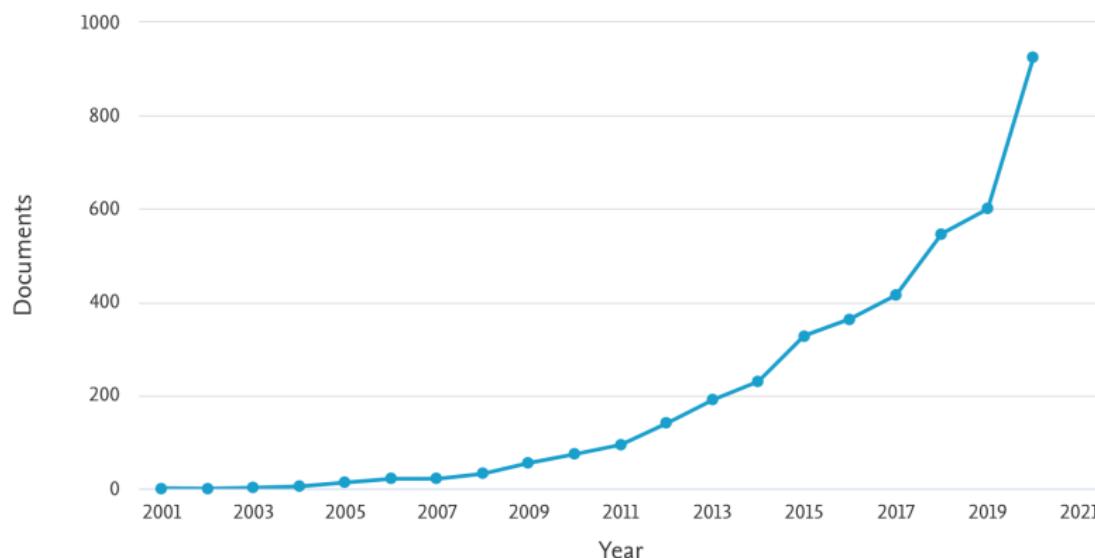
### ABSTRACT

In this study we introduce a new indicator for private consumption based on search query time series provided by Google Trends. The indicator is based on factors extracted from consumption-related search categories of the Google Trends application Insights for Search. The forecasting performance of the new indicator is assessed relative to the two most common survey-based indicators: the University of Michigan Consumer Sentiment Index and the Conference Board Consumer Confidence Index. The results show that in almost all conducted in-sample and out-of-sample forecasting experiments the Google indicator outperforms the survey-based indicators. This suggests that incorporating information from Google Trends may offer significant benefits to forecasters of private consumption. Copyright © 2011 John Wiley & Sons, Ltd.

KEY WORDS Google Trends; private consumption; forecasting; consumer senti-

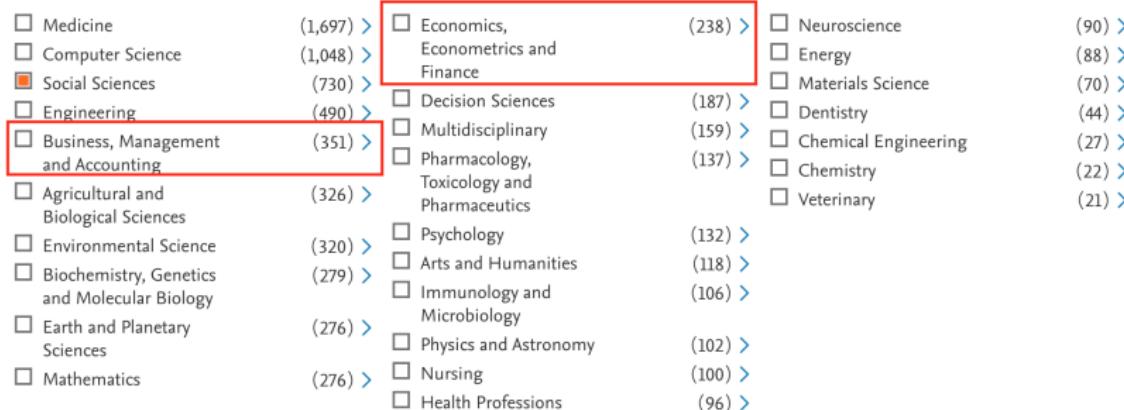
# Google Trends – nauka

Documents by year



Rysunek 3: Liczba artykułów naukowych indeksowanych w bazie Scopus zawierająca wyrażenie "Google Trends"

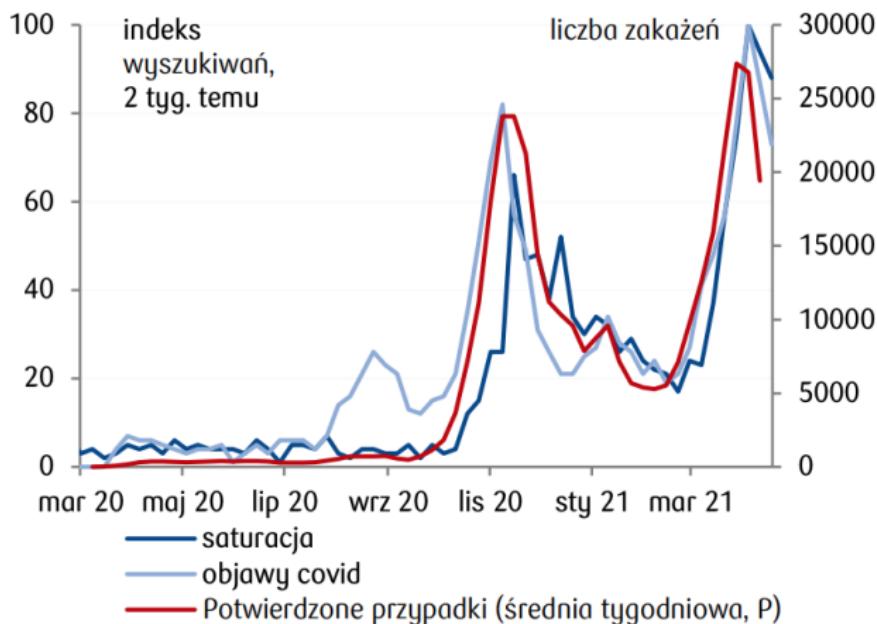
# Google Trends – nauka



Rysunek 4: Liczba artykułów naukowych indeksowanych w bazie Scopus zawierająca wyrażenie "Google Trends" według dziedziny

# Google Trends – COVID19

## Wyszukiwania związane z COVID-19



Rysunek 5: Zestawienie przygotowane przez zespół PKO PB Research

# Google Trends – podsumowanie

Ogólnie możemy podsumować korzystanie z Google Trends następująco:

- wykorzystanie indeksu Google Trends jako zmiennej w modelach progностycznych (prognozowanie krótkookresowe, nowcasting),
- wykorzystanie indeksu Google Trends w modelach przekrojowych / panelowych dodatkowe zmienne wyjaśniające dane zjawisko.

# Google Trends – jak korzystać?

Mamy dwie możliwości korzystania z Google Trends:

- oficjalna strona Google Trends:  
<https://trends.google.com/trends/?geo=PL>
- pakiety w językach programowania: `gtrendsR` (w R), `pytrends` (w Python).

# Google Trends – przykład

Przejdźmy do przykładu.

# Spis treści

- 1 O wykładowie
- 2 Internet w Polsce
- 3 Google Trends
- 4 Web scraping
- 5 Reprezentatywność
- 6 Metody estymacji dla prób nielosowych

# Czym jest web scraping?

- Jest to metoda wykorzystywanych do automatycznego pobierania i ekstrakcji zawartości stron stron internetowych (niezależnie od stosowanej technologii).
- W tym kontekście należy rozróżnić: *web crawling*, który polega na indeksowaniu stron/portali internetowych, a *web scraping*, który polega na ekstrakcji informacji z tych stron.
- Głównym elementem tej metody jest odpowiedni algorytm (tzw. *scraper*), który w małym (np. dopasowany do jednego portalu) lub dużym (np. rozpoznaje zwartość) stopniu działa automatycznie.
- Tożsame pojęcia: *Web scraping*, *web harvesting*, czy *web data extraction*
- Podstawowe technologie: WWW, html, przeglądarka internetowa.

# Czym jest web scraping? – idea

otodom.      Ogloszenia ▾      Rynek pierwotny ▾      Firmy ▾      Fixly ▾      Artykuły

Mieszkania ▾      na wynajem ▾      Poznań, wielkopolskie

Cena ▾      Obsługa zdalna HOME ▾      Powierzchnia ▾      Liczba pokoi ▾      Więcej filtrów ▾

Mieszkania na wynajem w Poznań, wielkopolskie      [Zapisz wyszukiwanie](#)

liczba ofert: 1 759      sortuj po: domyślnie ▾

[Promowane ogłoszenia](#)      [Zobacz wszystkie](#)

Najem lokalu bez pośrednictwa Czarnieckiego 6/96  
Mieszkanie na wynajem: Poznań, Wilda, Góra Wilda  
[Dodaj do ulubionych](#)

1 pokój 30,23 m<sup>2</sup>      1 299 zł /mc

Oferta prywatna

**Wynik web scrapingu**

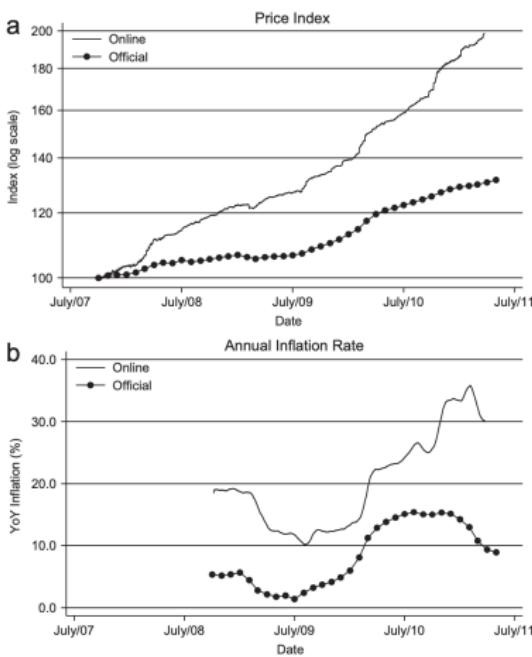
```
[1] "1800zł/mc" "1300zł/mc" "2750zł/mc" "1400zł/mc" "1700zł/mc"
[6] "1200zł/mc" "1650zł/mc" "1500zł/mc" "1400zł/mc" "1800zł/mc"
[11] "2400zł/mc" "2000zł/mc" "25000zł/mc" "2000zł/mc" "1100zł/mc"
[16] "1000zł/mc" "1400zł/mc" "3300zł/mc" "7500zł/mc" "1599zł/mc"
[21] "1100zł/mc" "1600zł/mc" "1300zł/mc" "2200zł/mc" "1850zł/mc"
[26] "1400zł/mc" "1600zł/mc"
```

# Web scraping w ekonomii – przykłady

## Wybrane publikacje:

- Edelman, B. (2012). Using internet data for economic research. *Journal of Economic Perspectives*, 26(2), 189-206.
- Cavallo, A. (2013). Online and official price indexes: Measuring Argentina's inflation. *Journal of Monetary Economics*, 60(2), 152-165.
- Cavallo, A., & Rigobon, R. (2016). The billion prices project: Using online prices for measurement and research. *Journal of Economic Perspectives*, 30(2), 151-78.
- Hoekstra, R., ten Bosch, O., & Hartevelde, F. (2012). Automated data collection from web sources for official statistics: First experiences. *Statistical Journal of the IAOS*, 28(3, 4), 99-111.
- Beręsewicz, M., Białkowska, G., Marcinkowski, K., Maślak, M., Opiela, P., Pater, R., & Zadroga, K. (2021). Enhancing the Demand for Labour survey by including skills from online job advertisements using model-assisted calibration. *Survey Research Methods (forthcoming)*
- Hołda, M. (2019). Newspaper-based economic uncertainty indices for Poland. *Narodowy Bank Polski*.

# Web scraping w ekonomii – przykłady



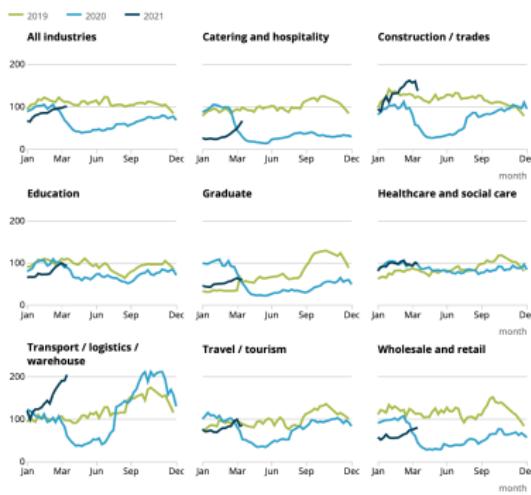
Rysunek 6: Źródło: Cavallo, A. (2013)

# Web scraping w ekonomii – przykłady

The screenshot shows the homepage of the European Centre for the Development of Vocational Training (CEDEFOP). The header features the CEDEFOP logo and name, along with a note about language availability. Below the header is a navigation menu with links to Home, Themes, Publications and resources, Events and projects, News and press, and Country Data. A yellow banner below the menu reads "CEDEFOP DATA VISUALISATIONS & TOOLS". Four circular icons represent different databases: "VET in Europe database" (map of Europe), "Financing apprenticeships database" (lightbulb and gears), "VET toolkit for tackling early leaving" (two people working together), and "Matching skills" (two people shaking hands). At the bottom, there's a "HEADLINES" section and a "View all" link, followed by a set of five small circular navigation dots.

Rysunek 7: Źródło: Cedefop

# Web scraping w ekonomii – przykłady



Source: Adzuna

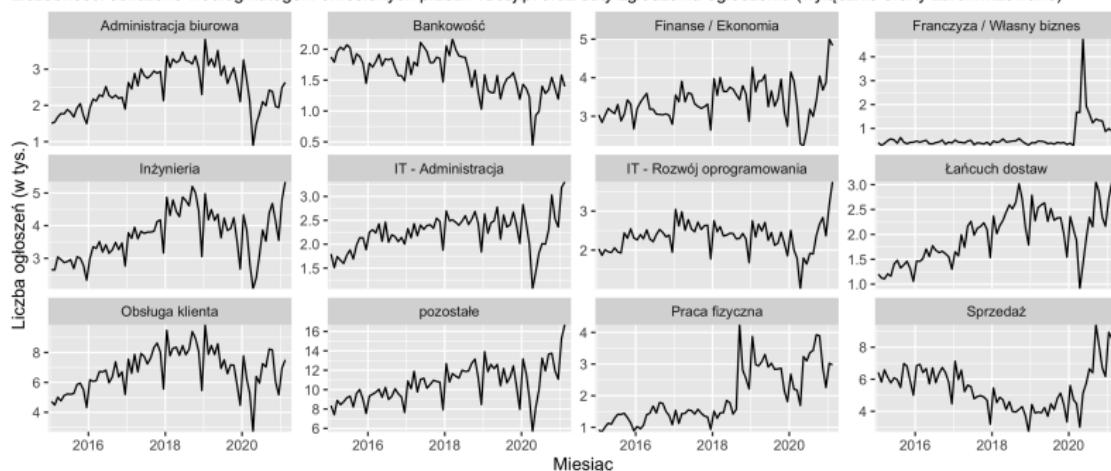
[Embed code](#)

Rysunek 8: Źródło: UK ONS – Coronavirus and the latest indicators for the UK economy and society: 22 April 2021

# Web scraping w ekonomii – przykłady

Liczba ogłoszeń na portalu Pracuj.pl w ujęciu miesięcznym w okresie 2015.01-2021.02

Licznebości obliczone według kategorii określonych przez Pracuj.pl oraz daty zgłoszenia ogłoszenia (wyłącznie oferty zarchiwizowane)



Źródło: archiwum.pracuj.pl. Liczba ogłoszeń: 3,3 mln. Dane częściowo wyczyszczone ale nie deduplikowane. Opracowanie: @mberesewicz.

Rysunek 9: Źródło: Opracowanie własne na podstawie pracuj.pl

# Web scraping w ekonomii – przykłady

Table 8: Point estimates of the fraction of skills for the pooled sample for 2011, 2013 and 2014

SKILLS	HTSRS	ECGREG	ECMC	ECLASSO1	ECLASSO2	ECALASSO1
Artistic	15.8	12.3	12.4	12.5	13.0	12.5
Availability	20.9	19.8	19.7	19.6	21.5	19.5
Cognitive	20.9	14.3	14.3	14.6	14.0	14.6
Computer	33.0	22.2	22.0	22.3	23.0	22.6
Interpersonal	53.8	34.5	34.5	35.1	35.0	34.9
Managerial	26.2	16.7	16.5	16.8	17.7	16.8
Mathematical	0.4	0.4	0.4	0.4	0.4	0.4
Office	3.9	3.1	3.1	3.2	3.4	3.2
Physical	5.4	7.4	7.6	7.5	8.2	7.6
Self-organization	58.6	43.8	43.5	43.9	46.2	43.8
Technical	4.3	7.5	7.7	7.7	8.3	7.7

Rysunek 10: Źródło: Beręsewicz et al. (2021)

# Web scraping w ekonomii – podsumowanie

Ogromna liczba zastosowań:

- mierniki ekonomii w czasie rzeczywistym (np. oferty pracy, nieruchomości, nastroje społeczne),
- weryfikacja różnych ekonomicznych konceptów (np. lepkość cen),
- now-casting / prognozowanie,
- rozszerzenie istniejących źródeł danych o informacje zawarte na portalach internetowych (np. popyt na pracę, RCiWN),
- zastąpienie wywiadów automatycznym pobieraniem informacji ze stron internetowych (np. pytanie o posiadanie social media, sklepu internetowego),
- aktualizacja operatów losowania (np. weryfikacja czy firma działa),
- wiele innych.

# Web scraping – wybrane technologie

Słowo wstępu:

- strony statyczne (HTML) vs dynamiczne (JavaScript),
- web scraping, a dostęp przez Application Programming Interface (API).

Przydatne biblioteki:

- Python – beautifulsoup, scrapy,
- R – rvest,
- Ogólne – PhantomJS.

# Web scraping – przykład

Przejdźmy do przykładu w R.

# Spis treści

1 O wykładzie

2 Internet w Polsce

3 Google Trends

4 Web scraping

5 Reprezentatywność

- Reprezentatywność – definicje, pomiar, problematyka

6 Metody estymacji dla prób nielosowych

## Zadanie – 5 minut

Której wersji bardziej Państwo zaufacie? Dlaczego?

- ➊ 1% próbie losowej z 60% realizacją (odsetkiem odpowiedzi), czy
- ➋ rejestrowi administracyjnemu tworzonemu przez dobrowolne deklaracje (*a self-reported administrative dataset*) pokrywającym 80% populacji?

## Zadanie – 5 minut

Co oznacza reprezentatywność? Jak to Państwo rozumiecie?

# Mocne prawo wielkich liczb

Dla ciągów (całkowalnych) zmiennych losowych wprowadza się definicję spełniania przez nich tzw. mocnego (i słabego) prawa wielkich liczb.

Ciąg zmiennych niezależnych zmiennych losowych  $(X_n)_{n=1}^{\infty}$  spełnia mocne prawo wielkich liczb (MPWL), gdy

$$\frac{1}{n} \sum_{k=1}^n (X_k - E(X_k)) \xrightarrow{p.n.} 0 \quad (1)$$

p.n. (Prawie na pewno): określenie zdarzenia zachodzącego z prawdopodobieństwem 1. Sformułowanie to pojawia się w naturalny sposób np. przy badaniu zagadnień granicznych.

# Estymator – własności

Niech  $\hat{\theta} = T(X_1, X_2, \dots, X_n)$  będzie estymatorem parametru  $\theta$ , to

- estymator jest nieobciążony gdy

$$E(\hat{\theta}) = \theta, \quad (2)$$

- obciążenie estymatora oznaczamy przez

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta, \quad (3)$$

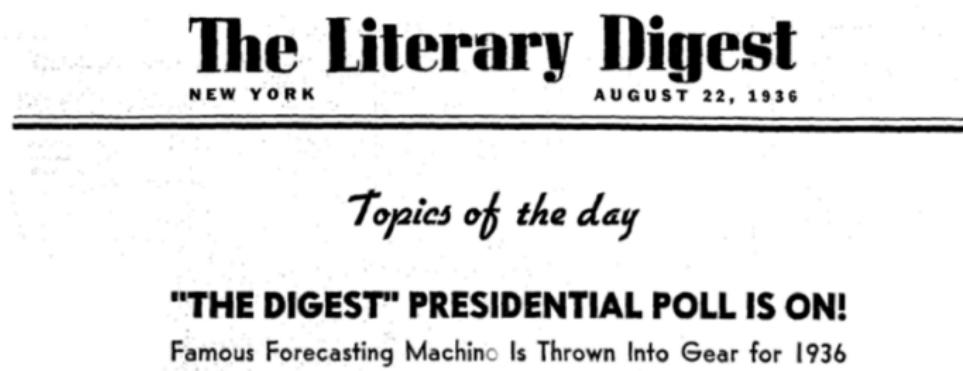
- wariancję estymatora natomiast

$$D^2(\hat{\theta}) = E(\hat{\theta} - E(\hat{\theta}))^2, \quad (4)$$

- estymator jest zgodny, gdy zachodzi

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| \leq \epsilon) = 1. \quad (5)$$

# Mała, a duża próba – The 1936 Literary Digest Poll



# Mała, a duża próba – The 1936 Literary Digest Poll

- The Literary Digest był bardzo wpływowym, amerykańskim tygodnikiem.
- Od 1916 roku poprawnie przewidywał wyniki wyborów w USA, a w **1936 roku odbywały się ważne wybory między Rooseveltem, a Landonem.**
- Tygodnik wysłał ponad **10 mln** karty do głosowania, a ponad **2.4 mln** kart zostało odesłanych.
  - *The mailing list was “drawn from every telephone book in the United States, from the rosters of clubs and associations, from city directories, lists of registered voters, classified mail-order and occupational data” (Lohr and Brick (2017))*
- Według tego badania wybory miał wygrać **Landon z 55%** poparciem, a **Roosevelt** miał otrzymać **41%**.
- W tym samym czasie Instytut Gallup'a wykorzystując próbę wielkości 50 tys. prognozował **56% dla Roosevelta**.
- Faktyczne wyniki były znaczco inne: **Roosevelt 61%, Landon 37%**.
- W 1938 roku Magazyn został zamknięty.

# The 1936 Literary Digest Poll – przyczyny

## The 1936 Literary Digest Poll

131

**Table 3.** Presidential Vote by Returning or Not Returning Straw Vote Ballot (in Percent)

Presidential Vote	Did Return	Did Not Return	Do Not Know
Roosevelt	48	69	56
Landon	51	30	40
Other	1	1	4
Total N	493	288	48

SOURCE: American Institute of Public Opinion, 28 May 1937.

**Rysunek 11:** Źródło: Squire, P. (1988). Why the 1936 Literary Digest poll failed. *Public Opinion Quarterly*, 52(1), 125-133.

# Przykłady

Reprezentatywność w opisie badań, na przykład:

- **Badanie EU-SILC** jest dobrowolnym, ‘reprezentacyjnym’ badaniem ankietowym prywatnych gospodarstw domowych, realizowanym techniką bezpośredniego wywiadu z respondentem.
- **Badanie Aktywności Ekonomicznej Ludności** przeprowadzane jest ‘metodą reprezentacyjną’, a wyniki badania uogólniane są na populację generalną. Z uwagi na reprezentacyjną metodę badania zalecana jest ostrożność w posługiwaniu się danymi w tych przypadkach, gdy zastosowano bardziej szczegółowe podziały i występują liczby niskiego rzędu (mniejsze niż 15 tys.).

# Reprezentatywność – definicja

Za słownikiem PWN (źródło: <http://sjp.pwn.pl/sjp/reprezentatywny;2515040.html>)

- **reprezentatywny** «mający cechy charakterystyczne dla jakiejś zbiorowości» (reprezentatywnie, reprezentatywność)

# Reprezentatywność – Próba reprezentatywna

## Próba reprezentatywna

- Próba, której struktura ze względu na badane cechy (zmiennne) jest zbliżona do struktury populacji statystycznej, z której pochodzi.
- Reprezentatywność próby (próba reprezentatywna) można uzyskać stosując zarówno losowe (probabilistyczne) jak i nielosowe (nieprobabilistyczne) techniki wyboru próby.
- Należy jednak zaznaczyć, iż większą szansę na reprezentatywność próby daje zastosowanie technik losowego jej wyboru.

Źródło: <https://stat.gov.pl/metainformacje/slownik-pojec/pojecia-s2771,pojcie.html>.

# Reprezentatywność – losowy dobór (za GUS)

## Losowy wybór próby

Technika pobierania próby z badanej populacji generalnej, która spełnia dwa następujące warunki:

- ① każda jednostka populacji ma dodatnie i znane prawdopodobieństwo dostania się do próby;
- ② dla każdego zespołu jednostek populacji można ustalić prawdopodobieństwo tego, że w całości znajdzie się on w próbie.

## Próba losowa

Próba pobrana za pomocą odpowiednich technik probabilistycznych (losowy wybór próby).

Źródło: <https://stat.gov.pl/metainformacje/slownik-pojec/pojecia-s2748,pojecie.html>.

# Reprezentatywność – nielosowy dobór (za GUS)

## Nielosowy wybór próby

Technika wyboru próby, która nie spełnia choć jednego z dwóch warunków określonych w definicji losowego wyboru próby.

Najpopularniejszymi technikami nielosowego wyboru próby są: wybór przypadkowy, wybór dogodny, wybór celowy i wybór kwotowy.

Źródło: <https://stat.gov.pl/metainformacje/slownik-pojec/pojecia-s-2761,pojecie.html>.

# Definicja reprezentatywności

Kruskal i Mosteller (1979a, 1979b, 1979c), na podstawie ówczesnej przeglądu literatury, podali następujące definicje reprezentatywności:

- ogólne, nieuzasadnione twierdzenie o danych (ang. *general, unjustified acclaim for the data*)
- losowy dobór do próby (ang. *absence of selective forces*)
- miniatura populacji (ang. *mirror or miniature of the population*)
- typowa jednostka (ang. *typical or ideal case(s)*)
- pokrycie populacji (ang. *coverage of the population*)
- termin bez uzasadnienia (ang. *a vague term to be made precise*)
- określona metoda doboru próby (ang. *representative sampling as a specific sampling method*)
- pozwala na nieobciążoną estymację (ang. *representative sampling as permitting good estimation*)
- dobór próby odpowiedni do danego problemu (ang. *representative sampling as good enough for a particular purpose*)

# Reprezentatywność – źródła danych

- ① **Źródła statystyczne** – źródła informacji statystycznej zaprojektowane przez statystyków i na potrzeby statystyki (np. badania częściowe, spisy powszechnie, rejestrystatystyczne, sprawozdawczość przedsiębiorstw)
- ② **Źródła niestatystyczne** – wszystkie pozostałe źródła danych, których celem nie jest dostarczanie informacji statystycznej o całej populacji (np. rejestrystad administracyjne, big data)

# Reprezentatywność – źródła danych wg Citro (2014)

Survey Methodology, December 2014  
 Vol. 40, No. 2, pp. 137-161  
 Statistics Canada, Catalogue No. 12-001-X

137

## From multiple modes for surveys to multiple data sources for estimates

Constance F. Citro<sup>1</sup>

### Abstract

Users, funders and providers of official statistics want estimates that are "wider, deeper, quicker, better, cheaper" (channeling Tim Holt, former head of the UK Office for National Statistics), to which I would add "more relevant" and "less burdensome". Since World War II, we have relied heavily on the probability sample survey as the best we could do - and that has been very good - to meet these goals for estimates of household income and unemployment, self-reported health status, time use, crime victimization, business activity, commodity flows, education and business outcomes, etc. Faced with secular declines in response rates and responses to evidence of sampling error, we have considered options including the use of multiple survey modes, more sophisticated weighting and imputation methods, adaptive design, cognitive testing of survey items, and other means to maintain data quality. For statistics on the business sector, in order to reduce burden and costs, we long ago moved away from relying solely on surveys to produce needed estimates, but, to date, we have not done that for household surveys, at least not in the United States. I argue that we can and must move from a paradigm of producing the best estimates possible from a survey to that of producing the best possible estimates to meet needs from multiple data sources. Such sources include administrative records and surveying, transaction and internet-based data. I provide two examples - household income and planning facilities - to illustrate my thesis. I suggest ways to inculcate a culture of official statistics that focuses on the end result of relevant, timely, accurate and cost-effective statistics and treats surveys, along with other data sources, as means to that end.

**Key Words:** Surveys; Administrative records; Total error; Big data; Income; Housing.

- **Constance Citro (ur. 1942) – amerykańska statystyczka i politolożka, m.in. była dyrektor *the Committee on National Statistics of the National Research Council*.**

**Table 5.1**  
**Ranking (HIGH, MEDIUM, LOW, VERY LOW, or VARIES) of four data sources on dimensions for use in official statistics**

Dimension/ Data Source	Census/Probability Survey (e.g., CPS/ASEC, ACS, NHIS - see Table 2.1)	Administrative Records (e.g., income taxes, Social Security, unemployment, payroll)	Commercial Transaction Records (e.g., scanner data, credit card data)	Individual Interactions with the Internet (e.g., Twitter postings; Google search term volumes)

## Przykłady – Selectivv

- Firma posiada obecnie największy w Europie Środkowo-Wschodniej zbiór informacji o właścicielach smartfonów i tabletów, **który obejmuje łącznie 82 mln osób, z czego 14 mln w Polsce.**
- Jesteśmy połączeni z sieciami reklamowymi, co daje nam dostęp do **200 tys. aplikacji i 15 milionów stron internetowych.**
- Średnio o jednym użytkowniku Selectivv **pozyskuje 362 informacje**, m.in. dane demograficzne, zainteresowania, jego styl życia oraz lokalizacje, w jakich przebywa. Wykorzystanie big data pozwoliło na wyróżnienie ponad 60 profili behawioralnych konsumentów, kategoryzując ich na m.in.: osoby planujące powiększenie rodziny, bywalców galerii handlowych, użytkowników bankowości mobilnej czy aplikacji muzycznych.

## Przykłady – banki

Liczba klientów indywidualnych			
Bank	III kw. 2020	II kw. 2020	III kw. 2019
PKO BP i Inteligo	<b>10 508 000</b>	10 465 700	10 401 000
Bank Pekao	<b>5 434 134</b>	5 388 766	5 349 673
Santander Bank Polska	<b>4 743 041</b>	4 698 385	4 610 781
Alior Bank i TMUB	<b>4 278 399</b>	4 211 010	4 075 953
ING Bank Śląski	<b>4 215 000</b>	4 133 000	4 288 000
mBank	<b>4 099 820</b>	4 086 000	3 979 263
Bank Millennium	<b>3 859 084</b>	3 810 561	2 693 843
BNP Paribas	<b>3 614 600</b>	3 626 700	3 500 000
Santander CB	<b>1 936 331</b>	1 977 819	2 091 847
Credit Agricole	<b>1 590 000</b>	1 620 000	1 725 047
Bank Pocztowy	<b>870 072</b>	898 358	930 541
BOŚ	<b>201 500</b>	202 700	223 900
Razem:	<b>45 349 981</b>	45 118 999	43 869 848

Rysunek 12: Liczba indywidualnych klientów banków w Polsce w III kw. 2020

Źródło: <https://prnews.pl/raport-prnews-pl-liczba-klientow-w-bank>

# Przykład – CBOP

 CENTRALNA BAZA OFERT PRACY  Unia Europejska

Język:  PL | Kontrast: A | Czcionka: A A+ A++ | Wsparcie:  Pomoc 

 Oferty pracy, staże i praktyki  Kalendarz targów, giełd i szkoleń  Wyszukiwanie pracowników  Zaloguj się  Zarejestruj się

Jesteś tutaj: [CBOP](#) > Oferty pracy, staże i praktyki

Liczba propozycji: 22 239, w tym w urzędach pracy **18 796** Ofert pracy **52 582** Wolnych miejsc pracy

 Wpisz nazwę stanowiska  Wpisz nazwę lokalizacji lub kod pocztowy + 0 km  Szukaj  
Wyszukiwanie zaawansowane

Sortowanie Data dodania  Poziom szczegółowości Niski  Pozycji na stronie 10  Strona 1 z 2224 następna >

WYBIERZ KRYTERIA	STANOWISKO	MIEJSCE PRACY	RODZAJ UMOWY	PRACODAIDCA	DOSTĘPNA OD
 WYBRANE KRYTERIUM	<a href="#">SPRZEDAJĄCA</a>	Bytom, śląskie	Umowa o pracę	kontakt przez PUP	dzisiaj

# Przykład – OtoDom

**otodom.** Ogłoszenia ▾ Rynek pierwotny ▾ Firmy ▾ Fixly ▾ Artykuły [Moje konto](#) [Dodaj ogłoszenie](#)

Mieszkania ▾ na sprzedaż ▾  [Wyszukaj](#)

Cena ▾ Powierzchnia ▾ Obsługa zdalna [NOWY](#) ▾ Liczba pokoi ▾ Rynek ▾Więcej filtrów ▾

Mieszkania na sprzedaż [Zapisz wyszukiwanie](#)

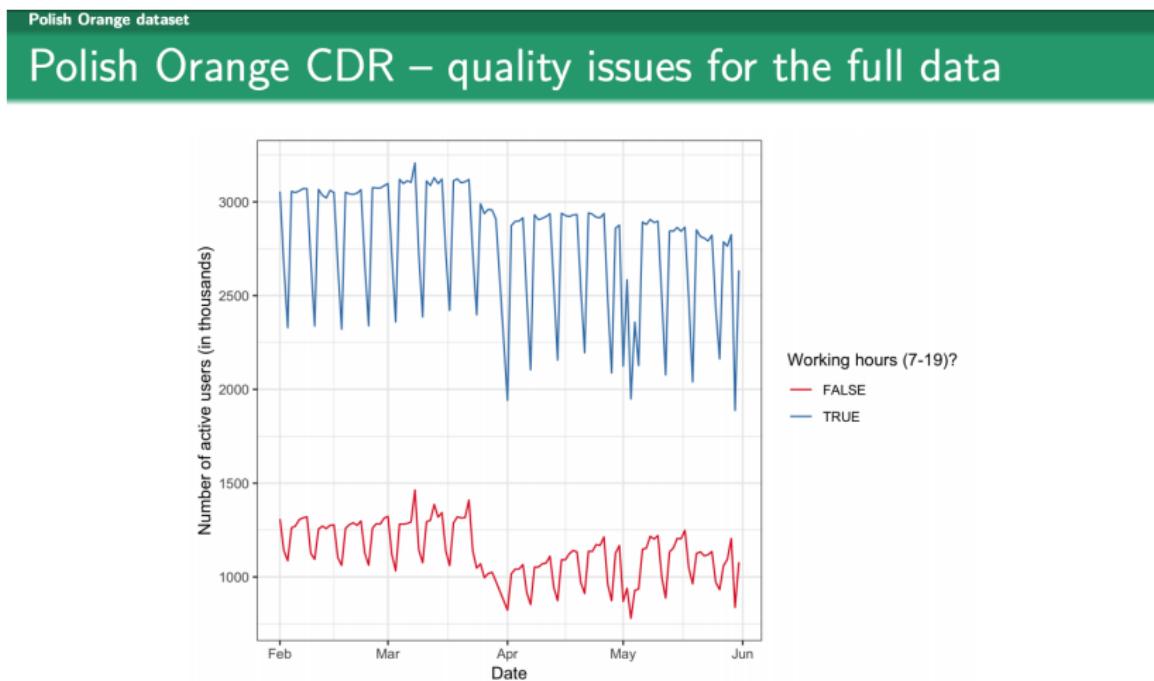
liczba ofert: **140 350**

Promowane ogłoszenia [Zobacz wszystkie](#)

sortuj po: domyślnie ▾ [Zobacz też oferty deweloperów](#)

 <b>3 Pokoje / Osiedle Leśne / M. W Hali / Premium</b> Mieszkanie na sprzedaż: Bydgoszcz, Osiedle Leśne <a href="#">Dodaj do ulubionych</a> 3 pokoje 77,67 m <sup>2</sup> 8 279 zł/m <sup>2</sup> <b>643 000 zł</b> Nasalski Nieruchomości i Finanse	 <b>Od 6 000 zł/m<sup>2</sup></b> <b>Szczecin, Zachód, Krzekowo DOM</b> Zamknięte osiedle domków na Krzekowie Nowe-domy.com Przedstawiciel dewelopera
 <b>Piękny, Nowy Dom Szeregowy, Balkon</b> Mieszkanie na sprzedaż: Wrocław, Psie Pole, Lipa Piotrowska <a href="#">Dodaj do ulubionych</a> 4 pokoje 65,60 m <sup>2</sup> 7 110 zł/m <sup>2</sup> <b>466 395 zł</b> Rydołoszcz, Górnny Taras.	 <b>Od 7 000 zł/m<sup>2</sup></b> <b>Bydgoszcz, Górnny Taras.</b>

# Przykład – Kałużny, Beręsewicz i Filipowska (2018)



**Figure 1:** Number of unique active users in working and non-working hours.  
Hours defined by the authors

Przejdźmy do precyzyjnych definicji reprezentatywności

# Podstawowe oznaczenia

- $I_i = \{0, 1\}$  – zmienna indykatrorowa; określa czy dana jednostka  $i$  była obserwowana w określonym zbiorze (np. ma dostęp do Internetu; ang. *inclusion*).
- $R_i = \{0, 1\}$  – zmienna indykatrorowa; określa czy dana jednostka  $i$  udzieliła odpowiedzi lub pseudo-odpowiedzi (np. umieściła wpis w Internecie; ang *response*).
- $Y$  – zmienna celu; cecha, którą badamy (np. poparcie dla danej partii)
- $y_i$  – wartość, którą obserwujemy dla zmiennej celu (np. wskazanie nazwy partii),
- $\mathbf{X}$  – zmienne pomocnicze; cechy, które uważamy, że są związane z  $Y$ ;  $\mathbf{x}_i$  – wartości cech  $\mathbf{X}$ , które obserwujemy w zbiorze danych.
- $P(I_i = 1)$  – prawdopodobieństwo pokrycia.
- $P(R_i = 1 | I_i == 1) = \rho_i$  – warunkowe prawdopodobieństwo pseudo-odpowiedzi.

# Silna i słaba reprezentatywność

Schouten i in. (2009) zaproponował dwie definicje reprezentatywności w odniesieniu do odpowiedzi (ang. *survey response*), które można ująć w kontekście losowego doboru do próby (braku autoselekcji), mianowicie:

- Silna reprezentatywność

$$\forall_i E(R_i) = \rho_i = P(R_i = 1 | I_i = 1) = \rho \quad (6)$$

- Słaba reprezentatywność

$$\bar{\rho}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} \rho_{ih} = \rho, \text{ for } h = 1, 2, \dots, H \quad (7)$$

Źródło: Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35(1), 101-113

# Miniatura populacji

- Próba jest reprezentatywna w odniesieniu do zmiennych pomocniczych ( $\mathbf{x}$ ) jeżeli rozkład cechy  $\mathbf{X}$  w próbie jest równy rozkładowi tej cechy w populacji przy założeniu losowego doboru próby

$$f_s(\mathbf{x}_i, l_i = 1) = f_{\Omega}(\mathbf{x}), \quad (8)$$

- Próba jest reprezentatywna w odniesieniu do zmiennych pomocniczych ( $\mathbf{x}$ ) jeżeli rozkład warunkowy cechy  $\mathbf{X}$  względem  $\mathbf{w}$  jest równy znanym wartościom globalnym (rozkładowi brzegowemu) tych cech w populacji generalnej.

$$f_s(\mathbf{x}_i | w_i, l_i = 1) = f_{\Omega}(\mathbf{x}), \quad (9)$$

gdzie  $w_i$  oznacza pewną wagę przypisaną danej jednostce  $i$ , którą może być zarówno odwrotność prawdopodobieństwa dostania się do próby  $w_i = d_i = 1/\pi_i$  czy wagi post-stratyfikowane czy kalibrowane.

# Reprezentatywny model – Pfeffermann (2011), Beręsewicz (2017)

Na podstawie Pfeffermann(2011) można wskazać pojęcie reprezentatywnego modelu, które może być zdefiniowane następująco:

Model jest reprezentatywny, wtedy i tylko wtedy gdy rozkład warunkowy cechy  $y$  pod warunkiem  $\mathbf{x}$  jest taki sam w próbie i populacji. To znaczy,  $f_s(y_i|\mathbf{x}_i) = f_\Omega(y_i|\mathbf{x}_i)$  tylko gdy

$$\Pr(R_i = 1 \mid \mathbf{x}_i, y_i, I_i = 1) = \Pr(R_i = 1 \mid \mathbf{x}_i, I_i = 1). \quad (10)$$

Pojęcie reprezentatywnego modelu możemy zapisać następująco:

$$f_s(y_i|\mathbf{x}_i) = f(y_i|\mathbf{x}_i, I_i = 1, R_i = 1) = \frac{\Pr(R_i = 1 \mid \mathbf{x}_i, y_i, I_i = 1) f_\Omega(y_i|\mathbf{x}_i)}{\Pr(R_i = 1 \mid \mathbf{x}_i, I_i = 1)}, \quad (11)$$

gdzie  $f_s(y_i|\mathbf{x}_i)$  jest rozkładem warunkowym w próbie,  $f_\Omega(y_i|\mathbf{x}_i)$  jest rozkładem warunkowym w populacji, a pozostałe elementy zdefiniowane są jak poprzednio.

## Betlehem (1988, 2002, 2010)

Betlehem (1988, 2002, 2010) pokazał, że obciążenie estymatora średniej z próby zdefiniowanej jako

$$\bar{y}_S = \frac{1}{n_S} \sum_{i=1}^N R_i Y_i, \quad (12)$$

której wartość oczekiwana w przypadku próby internetowej dana jest

$$E(\bar{y}_S) \approx \bar{Y}_I^* = \frac{1}{N_I \bar{\rho}} \sum_{i=1}^N \rho_k I_i Y_i, \quad (13)$$

może zostać zapisane jako

$$B(\bar{y}_S) = \frac{Corr(\rho, Y)\sigma_\rho\sigma_Y}{\bar{\rho}}, \quad (14)$$

gdzie  $Corr(\rho, Y)$  to korelacja między  $\rho$ , a  $Y$ ,  $\sigma$  to odchylenie standardowe, a  $\bar{\rho}$  to średnia arytmetyczna.

# Data Defect Index – Xiao-Li MENG (2018)

*The Annals of Applied Statistics*  
2018, Vol. 12, No. 2, 685–726  
<https://doi.org/10.1214/18-AOAS1161SF>  
© Institute of Mathematical Statistics, 2018

## STATISTICAL PARADISES AND PARADOXES IN BIG DATA (I): LAW OF LARGE POPULATIONS, BIG DATA PARADOX, AND THE 2016 US PRESIDENTIAL ELECTION<sup>1</sup>

BY XIAO-LI MENG

*Harvard University*

Statisticians are increasingly posed with thought-provoking and even paradoxical questions, challenging our qualifications for entering the statistical paradises created by Big Data. By developing measures for data quality, this article suggests a framework to address such a question: “Which one should I trust more: a 1% survey with 60% response rate or a self-reported administrative dataset covering 80% of the population?” A 5-element Euler-formula-like identity shows that for any dataset of size  $n$ , probabilistic or not, the difference between the sample average  $\bar{X}_n$  and the population average  $\bar{X}_N$  is the product of three terms: (1) a *data quality* measure,  $\rho_{R,X}$ , the correlation between  $X_j$  and the response/recording indicator  $R_j$ ; (2) a *data*

# Data Defect Index – Xiao-Li MENG (2018)

$$\overline{G}_n - \overline{G}_N = \underbrace{\rho_{R,G}}_{\text{Jakość danych}} \times \underbrace{\sqrt{\frac{1-f}{f}}}_{\text{Ilość danych}} \times \underbrace{\sigma_G}_{\text{Złożoność problemu}}, \quad (15)$$

gdzie:

- $G_j = G(X_j)$  – ogólna funkcja po zmiennych  $X$ , na przykład średnia arytmetyczna,
- $\overline{G}_n$  – średnia z próby o liczbeności  $n$ ,
- $\overline{G}_N$  – średnia w populacji o liczbeności  $N$ ,
- $\rho_{R,G}$  – korelacja między  $R$ , a wartościami  $G$  (**Uwaga:** tutaj  $\rho$  oznacza korelację  $\text{Corr}(R, G)$ , a nie  $\Pr(R = 1|X, Y)$  jak na wcześniejszych slajdach),
- $\sqrt{(1-f)/f} - f$  oznacza odsetek w próbie tj.  $f = n/N$ ,
- $\sigma_G$  – odchylenie standardowe dla wartości funkcji  $G$

Dowód: w artykule prof. Xiao-Li Menga.

# Prawo wielkich populacji (Meng, 2018)

W przypadku *Big data* i występowaniu autoselekcji mieżonej po następuje zmiana paradygmatu w ocenie błędów estymacji, tj.

przechodzimy z *prawa wielkich liczb* i *centralnego twierdzenia granicznego* według, którego

$$\text{error} \propto \frac{\sigma}{\sqrt{n}}, \quad (16)$$

do relatywnego systematycznego błędu (*prawa wielkich populacji*) według, którego

$$\text{error} \propto \hat{p}\sqrt{N}. \quad (17)$$

# Wybory w USA – Clinton vs Trump

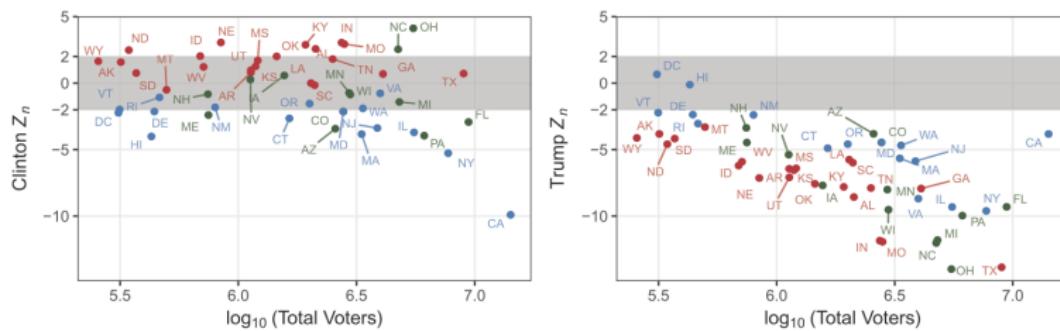


FIG. 7. Estimates of  $Z_n = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}}$ : The conventional 95% confidence interval region  $|Z_n| \leq 2$  is indicated in gray.

Rysunek 13: Obliczone na podstawie próby  $n = 2,315,570$  gdzie a  $N \approx 136,700,730$ . Meng oszacował, że  $\hat{p} = -0.005$ . Źródło: Meng (2018)

# DDI – jak bardzo mylimy się w przypadku dużych prób?

Celem przykładu będzie oszacowanie odsetka. Niech  $G_n()$  będzie odsetkiem 1 dla zmiennej  $Y = \{0, 1\}$ , stosując standardowy test proporcji (przy założeniu rozkładu normalności)  $Z_n$  będzie dany wzorem (za Mengiem)

$$Z_n = \frac{\hat{p} - p}{\sqrt{\hat{p}(1 - \hat{p})n}} = \frac{\sqrt{n}\sqrt{D_O}\rho_{R,G}}{\sqrt{1 - D_O\rho_{R,G}^2} - \sqrt{D_O}\rho_{R,G} \left( \sqrt{\frac{p}{1-p}} - \sqrt{\frac{1-p}{p}} \right)}, \quad (18)$$

gdzie  $D_O = (1 - f)/f$ . Założymy dla uproszczenia, że  $p = 0.5$  wtedy  $Z_n$  redukuje się do

$$Z_n = \sqrt{n} \sqrt{\frac{D_O\rho_{R,G}^2}{1 - D_O\rho_{R,G}^2}}. \quad (19)$$

Dzięki (19) możemy porównać co by było w pytaniu: co wolimy 1% próbę z 60% realizacją czy 80% próbę w postaci rejestru będącego wynikiem samo-rejestracji.

# DDI – jak bardzo mylimy się w przypadku dużych prób?

Założmy, że interesuje nas wnioskowanie o populacji Polski  $N=38\,000\,000$ .

**Tabela 1:** Porównanie  $Z_n$  dla próby 80% i 1%

Parametr	80% rejestr	1% próba
n	30,4 mln	380 tys.
$\rho_{R,G}$	0,005	0,001
$Z_n$	13.78	6.13

Liczba 13 i 6 oznacza, że mylimy się o odpowiednio 13 i 6 odchyleń standardowych. **Oznacza to, mając próbę 30 mln mylimy się bardziej niż mając losową próbę.**

*Uwaga:* przyjęliśmy tutaj, zgodnie z różnymi badaniami empirycznymi, że  $\rho_{R,G}$  jest zwykle większe dla prób nielosowych (przykładowo Meng (2018) szacował  $\hat{\rho}_{R,G} = -0.005$  dla poparcia dla Trumpa).

# Zależność między $X$ , $Y$ , a $R$

**Table 2**

*Effect of re-weighting (adapted from Table 1, Little & Vartivarian, 2005).*

	Low Association ( $X, Y$ )	High Association ( $X, Y$ )
Low association ( $X, R$ )	Little effect on bias; Little effect on variance	Little effect on bias; Variance reduction
High association ( $X, R$ )	Little effect on bias; Variance inflation	Bias reduction; Variance reduction

**Rysunek 14:** Zależność między cechami  $X$ ,  $Y$  oraz  $R$ .

Gdzie:  $X$  – zmienne pomocnicze,  $Y$  zmienna celu oraz  $R = \{0, 1\}$ .

Źródło: Zhang, L. C., Thomsen, I. B., & Kleven, Ø. (2013). On the Use of Auxiliary and Paradata for Dealing With Non-sampling Errors in Household Surveys. International Statistical Review, 81(2), 270-288.

# Jaki jest z tego wniosek?

- Redukcja obciążenia występuje tylko wtedy kiedy  $\mathbf{X}$  i  $Y$  są ze sobą silnie skorelowane ( $|Corr(Y, \mathbf{X})| > 0$ )
- Redukcja obciążenia występuje tylko wtedy kiedy  $\mathbf{X}$  i  $R$  są ze sobą silnie skorelowane ( $|Corr(R, \mathbf{X})| > 0$ )
- Korelacja między  $Y$  i  $R$  jest bliska zeru gdy ( $|Corr(Y, R|\mathbf{X})| \approx 0$ )
- Konieczne jest posiadanie informacji o  $X$ , a najlepiej gdybyśmy dysponowali zmienną  $X_k$ , która jest tzw. *proxy variable* – zbliżona ale nie ta sama definicja (np. cena ofertowa vs cena transakcyjna).

# Krótkie podsumowanie dotychczasowej wiedzy

Tabela 2: Próby losowe, a nielosowe

Czynnik	Próba losowa	Próba nielosowa
Dobór	Schemat losowania	Auto-selekcja
Pokrycie	Zwykle dobre	Pewne grupy są wykluczone
Obciążenie	Zwykle mniejsze	duże, lub bardzo duże
Wariancja	Zwykle większa	Mała, lub bardzo mała
Koszt	Duży lub bardzo duży	Zwykle nieduży

# Spis treści

1 O wykładzie

2 Internet w Polsce

3 Google Trends

4 Web scraping

5 Reprezentatywność

6 Metody estymacji dla prób nielosowych

- Podstawowe założenia na potrzeby zajęć
- Metody quasi-randomizacyjne
- Metody oparte na modelu
- Podwójnie odporne estymatory

# Metody estymacji – rozważmy następujące przypadki

A)

	X	Y
Próba losowa		
Próba nielosowa		

C)

	X	Y*	Y
Próba losowa			
Próba nielosowa			

B)

	X	Y
Próba losowa		
Próba nielosowa	Część wspólna obydwu prób	

D)

	X	Y
Rejestr / spis jednostek		
Próba nielosowa		

Rysunek 15: Cztery przykładowe przypadki źródeł danych, gdzie celem jest oszacowanie wybranej charakterystyki cechy  $Y$ . Cechy  $X$  są wspólne, cecha  $Y^*$  to tzw. zmienna proxy.

# Metody estymacji w przypadku prób nielosowych

*Statistical Science*  
2017, Vol. 32, No. 2, 249–264  
DOI: 10.1214/16-STS598  
© Institute of Mathematical Statistics, 2017

## Inference for Nonprobability Samples

Michael R. Elliott and Richard Valliant

*Abstract.* Although selecting a probability sample has been the standard for decades when making inferences from a sample to a finite population, incentives are increasing to use nonprobability samples. In a world of “big data”, large amounts of data are available that are faster and easier to collect than are probability samples. Design-based inference, in which the distribution for inference is generated by the random mechanism used by the sampler, cannot be used for nonprobability samples. One alternative is quasi-randomization in which pseudo-inclusion probabilities are estimated based on covariates available for samples and nonsample units. Another is superpopulation modeling for the analytic variables collected on the sample units in which the model is used to predict values for the nonsample units. We discuss the pros and cons of each approach.

*Key words and phrases:* Coverage error, hierarchical regression, quasi-randomization, reference sample, selection bias, superpopulation model.

# Elliott i Valliant (2017) wyróżniają dwa podejścia:

- **quasi-randomizacyjne** – w której konstrujemy *pseudo-wagi* z wykorzystaniem próby losowej lub znanych (albo estymowanych) wartości globalnych.

	X	W	Y	W*
Próba losowa				
Próba nielosowa				

Ostatecznie, do wnioskowania, korzystamy tylko z próby nielosowej

- **oparte na modelu** – w którym zakładamy pewien model.

	X	W	Y	
Próba losowa			$f(Y X) = Y_{pred}$	Ostatecznie, do wnioskowania, korzystamy tylko z próby losowej
Próba nielosowa				

# Podstawowe założenia

- Niech  $U = \{1, \dots, N\}$  oznacza populację docelową składającą się z  $N$  oznaczonych jednostek. Każda jednostka  $i$  ma powiązany wektor zmiennych pomocniczych  $\mathbf{x}_i$  oraz zmienną badaną  $y_i$ .
  - $\{(y_i, \mathbf{x}_i), i \in S_A\}$  - próba nielosowa  $S_A$  o liczbie  $n_A$
  - $\{(\mathbf{x}_i, \pi_i^B), i \in S_B\}$  - próba losowa  $S_B$  o liczbie  $n_B$
- Dla próby losowej  $S_B$  każda jednostka ma przypisaną wagę wynikającą z schematu losowania:  $d_i^B = 1/\pi_i^B$ , gdzie  $\pi_i^B$  to prawdopodobieństwo włączenia.
- Naszym celem jest oszacowanie średniej w populacji  $\mu_y = N^{-1} \sum_{i=1}^N y_i$  zmiennej  $y$ .

# Wskaźniki i prawdopodobieństwa włączenia

- Wprowadźmy wskaźniki włączenia do prób, zdefiniowane dla wszystkich jednostek w populacji docelowej:
  - $R_i^A = I(i \in S_A)$  - włączenie do próby nielosowej  $S_A$
  - $R_i^B = I(i \in S_B)$  - włączenie do próby losowej  $S_B$
- Niech  $\pi_i^A = P(R_i^A = 1 | \mathbf{x}_i, y_i) = P(R_i^A = 1 | \mathbf{x}_i)$  będzie prawdopodobieństwem włączenia (propensity score), które charakteryzuje mechanizm doboru próby  $S_A$ .
- W przeciwnieństwie do  $\pi_i^B$ , wartości  $\pi_i^A$  i odpowiadające im wagi  $d_i^A = 1/\pi_i^A$  są nieznane.

# Struktura danych

Próba	ID	Włączenie ( $R$ )	Waga ( $d$ )	Zmienne pom. ( $x$ )	Zm. badana ( $y$ )
Nielosowa	1	1	?	✓	✓
$S_A$	:	:	:	:	:
	$n_A$	1	?	✓	✓
Losowa	1	0	✓	✓	?
$S_B$	:	:	:	:	:
	$n_B$	0	✓	✓	?

Tabela 3: Struktura danych w układzie dwóch prób

Wartości zmiennej  $y_i$  nie są obserwowane w próbie losowej, więc nie można ich bezpośrednio wykorzystać do oszacowania badanej wielkości. Zamiast tego, próbujemy łączyć próbę nielosową i losową.

## Główne założenia

Ze względu na brak uniwersalnie przyjętej metody łączenia prób, założenia różnią się znacznie. Jednak dla większości metod przyjmuje się następujące główne założenia:

- A1  $R_i^A$  i zmienna badana  $y_i$  są niezależne przy ustalonym zbiorze zmiennych pomocniczych  $\mathbf{x}_i$  (mechanizm MAR).
- A2 Wszystkie jednostki w populacji docelowej mają niezerowe prawdopodobieństwo włączenia do próby nielosowej, tj.  $\pi_i^A > 0$ ,  $i = 1, 2, \dots, N$  (brak błędu pokrycia).
- A3 Wskaźniki włączenia  $R_1, R_2, \dots, R_N$  są niezależne przy ustalonym zbiorze zmiennych pomocniczych  $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$  (brak grupowania).

Obecnie pomijamy nakładanie się prób  $S_A$  i  $S_B$  oraz zakładamy brak błędów pomiaru w  $y_i$  i znajomość wartości  $\mathbf{x}_i$ .

# Propensity score – Rosenbaum & Rubin (1983)

*Biometrika* (1983), **70**, 1, pp. 41–55

Printed in Great Britain

41

## The central role of the propensity score in observational studies for causal effects

BY PAUL R. ROSENBAUM

*Departments of Statistics and Human Oncology, University of Wisconsin, Madison,  
Wisconsin, U.S.A.*

AND DONALD B. RUBIN

*University of Chicago, Chicago, Illinois, U.S.A.*

### SUMMARY

The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates. Both large and small sample theory show that adjustment for the scalar propensity score is sufficient to remove bias due to all observed covariates. Applications include: (i) matched sampling on the univariate propensity score, which is a generalization of discriminant matching, (ii) multivariate adjustment by subclassification on the propensity score where the same subclasses are used to estimate treatment effects for all outcome variables and in all subpopulations, and (iii) visual representation of multivariate covariance adjustment by a two-

# Propensity score – Lee (2006)

Journal of Official Statistics, Vol. 22, No. 2, 2006, pp. 329–349

## Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys

*Sunghee Lee<sup>1</sup>*

Propensity score adjustment (PSA) has been suggested as an approach to adjustment for volunteer panel web survey data. PSA attempts to decrease, if not remove, the biases arising from noncoverage, nonprobability sampling, and nonresponse in volunteer panel web surveys. Although PSA is an appealing method, its application in web survey practice is not well documented, and its effectiveness is not well understood. This study attempts to provide an overview of the PSA application by demystifying its performance for web surveys. Findings are three-fold: (a) PSA decreases bias but increases variance, (b) it is critical to include covariates that are highly related to the study outcomes, and (c) the role of nondemographic variables does not seem critical to improving PSA.

*Key words:* Web survey; propensity score adjustment

# Propensity score – Chen (2020)

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION  
2020, VOL. 115, NO. 532, 2011–2021: Theory and Methods  
<https://doi.org/10.1080/01621459.2019.1677241>



## Doubly Robust Inference With Nonprobability Survey Samples

Yilin Chen, Pengfei Li, and Changbao Wu

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada

### ABSTRACT

We establish a general framework for statistical inferences with nonprobability survey samples when relevant auxiliary information is available from a probability survey sample. We develop a rigorous procedure for estimating the propensity scores for units in the nonprobability sample, and construct doubly robust estimators for the finite population mean. Variance estimation is discussed under the proposed framework. Results from simulation studies show the robustness and the efficiency of our proposed estimators as compared to existing methods. The proposed method is used to analyze a nonprobability survey sample collected by the Pew Research Center with auxiliary information from the Behavioral Risk Factor Surveillance System and the Current Population Survey. Our results illustrate a general approach to inference with nonprobability samples and highlight the importance and usefulness of auxiliary information from probability survey samples. Supplementary materials for this article are available online.

### ARTICLE HISTORY

Received May 2018  
Accepted September 2019

### KEYWORDS

Design-based inference;  
Inclusion probability; Missing  
at random; Propensity score;  
Regression modeling;  
Variance estimation

# Propensity score – Wu (2022)

Survey Methodology, December 2022  
Vol. 48, No. 2, pp. 283-311  
Statistics Canada, Catalogue No. 12-001-X

283

## Statistical inference with non-probability survey samples

Changbao Wu<sup>1</sup>

### Abstract

We provide a critical review and some extended discussions on theoretical and practical issues with analysis of non-probability survey samples. We attempt to present rigorous inferential frameworks and valid statistical procedures under commonly used assumptions, and address issues on the justification and verification of assumptions in practical applications. Some current methodological developments are showcased, and problems which require further investigation are mentioned. While the focus of the paper is on non-probability samples, the essential role of probability survey samples with rich and relevant information on auxiliary variables is highlighted.

**Key Words:** Auxiliary information; Bootstrap variance estimator; Calibration method; Doubly robust estimator; Estimating equations; Inverse probability weighting; Model-based prediction; Poststratification; Pseudo likelihood; Propensity score; Quota survey; Sensitivity analysis; Variance estimation.

# Nazewnictwo

- Propensity score (PS) – prawdopodobieństwa inkluzji do próby nielosowej  $S_A$
- Propensity score weighting (PSW) / Inverse probability weighting (IPW) – ważenie przez odwrotność prawdopodobieństwa inkzji

# Wprowadzenie do metody IPW

- Inverse Probability Weighting (IPW) to popularna metoda estymacji wykorzystująca prawdopodobieństwa włączenia do próby
- Opiera się na estymacji prawdopodobieństwa włączenia (PS) danego wzorem  $\pi_i^A = P(i \in S_A)$
- Metoda szczególnie przydatna w kontekście próby nielosowej
- Prawdopodobieństwa włączenia są używane do korygowania obciążień wynikających z mechanizmu selekcji próby

## Dwa warianty estymatora IPW

Metoda IPW oferuje dwa główne warianty estymatora średniej populacyjnej:

$$\hat{\mu}_{y, \text{IPW1}} = \frac{1}{N} \sum_{i \in S_A} \frac{y_i}{\hat{\pi}_i^A} \quad \text{ i } \quad \hat{\mu}_{y, \text{IPW2}} = \frac{1}{\hat{N}^A} \sum_{i \in S_A} \frac{y_i}{\hat{\pi}_i^A}, \quad (20)$$

gdzie:

- $\hat{\mu}_{y, \text{IPW1}}$  - modyfikacja estymatora Horwitza-Thompsona
- $\hat{\mu}_{y, \text{IPW2}}$  - modyfikacja estymatora Hájeka
- $\hat{N}^A = \sum_{i \in S_A} (\hat{\pi}_i^A)^{-1}$  - oszacowana liczebność populacji

Wu (2022) zauważa, że estymator  $\hat{\mu}_{y, \text{IPW2}}$  daje lepsze wyniki niż  $\hat{\mu}_{y, \text{IPW1}}$  nawet gdy liczebność populacji jest znana.

# Konstrukcja estymatora IPW

Budowa estymatora IPW wymaga dwóch kroków:

- ➊ Estymacja prawdopodobieństw włączenia  $\pi_i^A$
- ➋ Wyznaczenie wag  $d_i^A = 1/\pi_i^A$

Do estymacji prawdopodobieństw  $\pi_i^A = \pi(\mathbf{x}_i, \gamma)$  można zastosować metodę największej wiarygodności, zakładając, że informacje o  $\mathbf{x}_i$  są dostępne dla każdej jednostki w populacji (lub próby losowej).

# Regresja logistyczna w estymacji PS

- W praktyce, do modelowania prawdopodobieństwa włączenia  $\pi(\mathbf{x}_i, \gamma)$  najczęściej stosuje się regresję logistyczną
- Model logistyczny przyjmuje postać:

$$\pi(\mathbf{x}_i, \gamma) = \frac{\exp(\mathbf{x}_i^T \gamma)}{1 + \exp(\mathbf{x}_i^T \gamma)} \quad (21)$$

- Parametry  $\gamma$  mają interpretację logarytmu ilorazu szans
- Model jest elastyczny i może uwzględniać zmienne ilościowe, jakościowe oraz interakcje

# Estymacja modelu logistycznego

- Do estymacji parametrów  $\gamma$  modelu logistycznego używamy próby nie-losowej ( $S_A$ ) i losowej ( $S_B$ )
- Jednostki z próby nielosowej traktujemy jako "sukcesy" ( $R_i^A = 1$ )
- Jednostki z próby losowej traktujemy jako "porażki" ( $R_i^A = 0$ )
- Zmienne objaśniające  $x_i$  powinny dobrze przewidywać mechanizmłączenia do próby nielosowej
- Ważne jest uwzględnienie wag  $d_i^B$  z próby losowej w procesie estymacji

**Uwaga:** Jakość oszacowań  $\hat{\pi}_i^A$  bezpośrednio wpływa na jakość finalnego estymatora IPW. Istotny jest dobór odpowiednich zmiennych pomocniczych  $x_i$  silnie związanych zarówno z mechanizmemłączenia do próby jak i zmienną badaną  $y$ .

# Warunki stosowania modelu logistycznego

- Wymaga spełnienia założenia MAR (Missing At Random) - prawdopodobieństwo włączenia do próby zależy tylko od obserwowań zmiennych  $x_i$ ,
- Wymaga odpowiedniego pokrycia przestrzeni zmiennych pomocniczych w próbie losowej i nielosowej
- Możliwe jest rozszerzenie do bardziej złożonych modeli (np. modele GAM, lasy losowe) jeśli relacja nie jest liniowa w logice

Po estymacji modelu, otrzymane prawdopodobieństwa  $\hat{\pi}_i^A$  są używane do konstruowania wag  $\hat{d}_i^A = 1/\hat{\pi}_i^A$  stosowanych w estymatorach IPW.

# Funkcja wiarygodności

Teoretyczna funkcja logarytmu wiarygodności ma postać:

$$\begin{aligned}\ell(\gamma) &= \log \left\{ \prod_{i=1}^N \left( \pi_i^A \right)^{R_i} \left( 1 - \pi_i^A \right)^{1-R_i} \right\} \\ &= \sum_{i \in S_A} \log \left\{ \frac{\pi(x_i, \gamma)}{1 - \pi(x_i, \gamma)} \right\} + \sum_{i=1}^N \log \{1 - \pi(x_i, \gamma)\}. \quad (22)\end{aligned}$$

W praktyce funkcja tej postaci nie może być stosowana, ponieważ nie wszystkie jednostki z populacji są obserwowane.

# Pseudo funkcja wiarygodności

Bardziej realistyczne podejście polega na wykorzystaniu referencyjnej próby losowej  $S_B$ . Wtedy otrzymujemy pseudo funkcję logarytmu wiarygodności:

$$\ell^*(\gamma) = \sum_{i \in S_A} \log \left\{ \frac{\pi(\mathbf{x}_i, \gamma)}{1 - \pi(\mathbf{x}_i, \gamma)} \right\} + \sum_{i \in S_B} d_i^B \log \{1 - \pi(\mathbf{x}_i, \gamma)\}. \quad (23)$$

Estymator  $\hat{\gamma}$  największej pseudo-wiarygodności można uzyskać jako rozwiązanie równania pseudo-score, które przy założeniu funkcji logitowej dla  $\pi_i^A$  ma postać:

$$\mathbf{U}(\gamma) = \sum_{i \in S_A} \mathbf{x}_i - \sum_{i \in S_B} d_i^B \pi(\mathbf{x}_i, \gamma) \mathbf{x}_i. \quad (24)$$

Przejdźmy do przykładu w R

## Ograniczenia IPW MLE

Estymacja  $\pi(\mathbf{x}_i, \gamma)$  z wykorzystaniem metody MLE lub pseudo-MLE ma następujące ograniczenia

- działa tylko jeżeli model dla  $\pi(\mathbf{x}_i, \gamma)$  jest poprawnie wyspecyfikowany,
- może skutkować dużymi wagami gdy  $\pi(\mathbf{x}_i, \gamma)$  jest małe,
- wariancja wag zwykle jest większa,
- wagi  $1/\pi(\mathbf{x}_i, \gamma)$  nie odtwarzają wartości globalnych z próby lub populacji,
- możemy zastosować tylko gdy dysponujemy danymi z populacji lub próby.

# Rozszerzenia i metoda kalibracji

Funkcja  $\mathbf{U}(\gamma)$  może być zastąpiona przez ogólne równania estymujące (ang. *generalized estimation equations*; GEE).

- Niech  $\mathbf{h}(\mathbf{x}, \gamma)$  będzie wektorem funkcji o tym samym wymiarze co  $\gamma$
- Definiujemy funkcję  $\mathbf{G}(\gamma)$  jako:

$$\mathbf{G}(\gamma) = \sum_{i \in S_A} \mathbf{h}(\mathbf{x}_i, \gamma) - \sum_{i \in S_B} d_i^B \pi(\mathbf{x}_i, \gamma) \mathbf{h}(\mathbf{x}_i, \gamma) \quad (25)$$

- Rozwiążanie równania  $\mathbf{G}(\gamma) = \mathbf{0}$  daje zgodny estymator  $\hat{\gamma}$
- Najczęściej wybierane funkcje to
  - $\mathbf{h}(\mathbf{x}_i, \gamma) = \mathbf{x}_i \pi(\mathbf{x}_i, \gamma)^{-1}$ ,
  - $\mathbf{h}(\mathbf{x}_i, \gamma) = \mathbf{x}_i$  (prowadzi do  $\mathbf{U}(\gamma)$ ).

Dla drugiego wariantu funkcji  $\mathbf{h}$  otrzymujemy *skalibrowane IPW*:

$$\mathbf{G}(\theta) = \sum_{i \in S_A} \frac{\mathbf{x}_i}{\pi(\mathbf{x}_i, \gamma)} - \sum_{i \in S_B} d_i^B \mathbf{x}_i. \quad (26)$$

## Rozszerzenia i metoda kalibracji

Proszę zauważyć, że dla (26) sumę daną wzorem  $\sum_{i \in S_B} d_i^B \mathbf{x}_i$  możemy zastąpić wartościami globalnym (znanymi z populacji)

$$\mathbf{G}(\boldsymbol{\theta}) = \sum_{i \in S_A} \frac{\mathbf{x}_i}{\pi(\mathbf{x}_i, \boldsymbol{\gamma})} - \boldsymbol{\tau}_{\mathbf{x}} = \begin{bmatrix} \sum_{i \in S_A} \frac{\mathbf{x}_{i1}}{\pi(\mathbf{x}_i, \boldsymbol{\gamma})} \\ \sum_{i \in S_A} \frac{\mathbf{x}_{i2}}{\pi(\mathbf{x}_i, \boldsymbol{\gamma})} \\ \dots \\ \sum_{i \in S_A} \frac{\mathbf{x}_{ip}}{\pi(\mathbf{x}_i, \boldsymbol{\gamma})} \end{bmatrix} - \begin{bmatrix} \tau_{x_1} \\ \tau_{x_2} \\ \dots \\ \tau_{x_p} \end{bmatrix} \quad (27)$$

Kim i Riddles (2012) pokazali, że gdy estymujemy  $\boldsymbol{\gamma}$  z wykorzystaniem GEE to estymator IPW jest podwójnie odporny przy założeniu, że model zmiennej wynikowej ( $Y$ ) jest liniowy (tj. liniowy w populacji).

# Rozszerzenia i metoda kalibracji

Metoda GEE ma następujące własności:

- estymator ten jest podwójnie odporny i.e. kalibrując do znanych wartości globalnych zakładamy, że model  $E(Y|X)$  jest liniowy, por. Kim and Riddles (2012). *Some theory for propensity scoring adjustment estimator*, Survey Methodology 38, 157-165 – oznacza to, że jeżeli model dla  $\pi(x_i, \gamma)$  będzie źle wyspecyfikowany ale model  $E(Y|X)$  jest liniowy to estymator dostarczy nam (asymptotycznie) nieobciążonych szacunków.
- wagi  $1/\pi(x_i, \gamma)$  odtwarzają wartości globalne,
- estymator GEE możemy stosować gdy nie mamy dostępu do danych jednostkowych a jedynie wartości globalnych,
- wariancja estymatora GEE jest zwykle mniejsza niż IPW,
- wariancja wag  $1/\pi(x_i, \gamma)$  w przypadku metody GEE jest zwykle mniejsza i zwykle unikamy wag ekstremalnych.

# Estymacja wariancji IPW

- Chen et al. (2020) przedstawili postać analityczną oraz metodę bootstrap umożliwiającą estymację wariancji estymatora IPW dla średniej.
- W pakiecie `nonprobsvy` zaimplementowaliśmy zarówno podejście analityczne, jak i metodę bootstrap.
- Idea metody bootstrap polega na,  $b = 1, \dots, B$  razy powtarzamy
  - Losujemy ze zwracaniem jednostki z próby  $S_A$  (nielosowej)
  - Losujemy ze zwracaniem jednostki z próby  $S_B$  (losowej) zgodnie ze schematem losowania.
  - Estymujemy parametry  $\gamma$  funkcji propensity score  $\pi(\mathbf{x}_i, \gamma)$  i wyznaczamy  $\hat{\mu}_{y, \text{IPW}2,b}$

Następnie wyznaczamy wariancję z wektora oszacowań

$$(\hat{\mu}_{y, \text{IPW}2,1}, \hat{\mu}_{y, \text{IPW}2,2}, \dots, \hat{\mu}_{y, \text{IPW}2,b})^T.$$

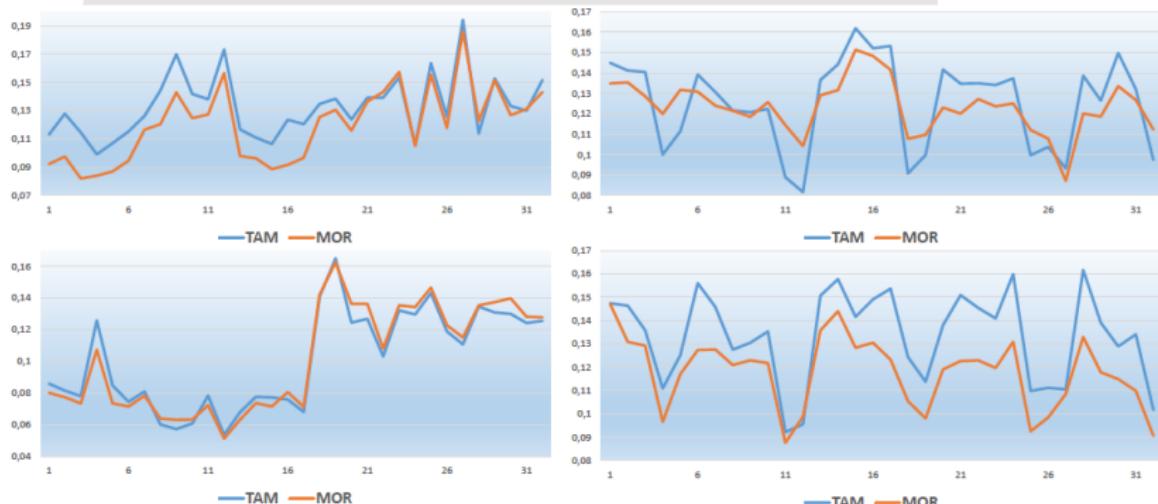
Przejdźmy do przykładu w R

## Przykład – pomiar oglądalności

- Oglądalność telewizji mierzona jest w badaniu *Nielsen Audience Measurement* (NAM) – próba losowa 2,5 tys. gospodarstw domowych (ok. 9,7 tys. osób)
- Alternatywa: model oglądalności rzeczywistej (MOR) od Netii – 180 tys. klientów (gospodarstw domowych) Netii
- Sposób ważenia: liczba gospodarstw i osób w województwie i wielkości miejsca zamieszkania (zgodnie z informacjami z 2018 roku).
- Na podstawie wyników NAM wycenia się reklamy w mediach oraz dochodzi do rozliczeń domów mediowych z nadawcami.

# Przykład – pomiar oglądalności

## Przykładowe udziały – TAM (|) vs MOR (||)



Rysunek 16: Porównanie badania NAM i Netii. Źródło: Wirtualne Media

# Przykład – pomiar oglądalności

## Krytyka badania Netii

- Zakłada się, że danym momencie wszystkie osoby oglądają telewizję.
- Nie wiadomo, kto ogląda i czy w ogóle ogląda telewizję.
- Nie jest znana liczba gospodarstw domowych w Polsce – to jest estymowane na podstawie danych z NSP czy badań częściowych.
- Ważenie dokonuje się wyłącznie na podstawie danych z województw i miejsca zamieszkania zakładając średnią liczbę osób w gospodarstwie domowym (wg. NSP 2021 była to średnio 2,99 na gospodarstwo domowe w Polsce).

# Przykład – pomiar oglądalności

Źródła:

- Jak mierzy się oglądalność telewizji w Polsce?  
<https://www.wirtualnemedia.pl/artykul/ogladowosc-telewizji-w-polsce-jak-to-sie-mierzy-nielsen>
- Telewizja Polska podaje szczegóły własnego badania oglądalności z danymi od Netii. Jacek Kurski: to projekt przejściowy  
<https://www.wirtualnemedia.pl/artykul/telewizja-polska-nowe-badanie-ogladowosci-z-danymi-od-netii>
- Domy mediowe: badanie oglądalności od TVP niemiarodajne, nie zastąpi pomiaru Nielsena  
<https://www.wirtualnemedia.pl/artykul/nowe-badanie-ogladowosci-od-tvp-i-netii-domy-mediowe-jest->

# Model semi-parametryczny w podejściu predykcyjnym

- W podejściu predykcyjnym zakłada się semi-parametryczny model dla populacji skończonej:

$$E_{\xi}(y_i | \mathbf{x}_i) = m(\mathbf{x}_i, \boldsymbol{\beta}), \quad V_{\xi}(y_i | \mathbf{x}_i) = v(\mathbf{x}_i) \sigma^2$$

- Funkcje  $m(\cdot, \cdot)$  i  $v(\cdot)$  mają znane formy
- $y_i$  są warunkowo niezależne, gdy dane są  $\mathbf{x}_i$
- Model obowiązuje dla wszystkich jednostek w próbie nieprobabilistycznej  $S_A$
- Parametry modelu można oszacować metodą quasi-największej wiarygodności

# Estymatory predykcyjne

- Dwa powszechnie stosowane estymatory predykcyjne:

$$\hat{\mu}_{y,PR1} = \frac{1}{N} \sum_{i=1}^N \hat{m}_i \quad i \quad \hat{\mu}_{y,PR2} = \frac{1}{N} \left\{ \sum_{i \in S_A} y_i - \sum_{i \in S_A} \hat{m}_i + \sum_{i=1}^N \hat{m}_i \right\}$$

- Dla modeli liniowych, gdzie  $m(\mathbf{x}_i, \beta) = \mathbf{x}'_i \beta$ , estymatory upraszczają się do:

$$\hat{\mu}_{y,PR1} = \boldsymbol{\mu}'_x \hat{\beta} \quad i \quad \hat{\mu}_{y,PR2} = \frac{n_A}{N} \left( \bar{y}_A - \bar{\mathbf{x}}'_A \hat{\beta} \right) + \boldsymbol{\mu}'_x \hat{\beta}$$

- Jeśli model liniowy zawiera wyraz wolny i  $\hat{\beta}$  jest estymatorem najmniejszych kwadratów, to  $\hat{\mu}_{y,PR1} = \hat{\mu}_{y,PR2}$

# Korzyści z estymatorów predykcyjnych

- Formuła  $\hat{\mu}_{y,PR2}$  wymaga tylko:
  - próby nieprobabilistycznej  $S_A$
  - średnich populacyjnych (lub sum i wielkości populacji  $N$ )
- Jeśli średnie populacyjne są nieznane, można je zastąpić estymatorami z referencyjnej próby probabilistycznej  $S_B$ :
  - $\sum_{i=1}^N \hat{m}_i$  zastępuje się przez  $\sum_{i \in S_B} d_i^B \hat{m}_i$
  - $\mu_x$  zastępuje się przez  $\hat{\mu}_x = \hat{N}_B^{-1} \sum_{i \in S_B} d_i^B x_i$ , gdzie  $\hat{N}_B = \sum_{i \in S_B} d_i^B$

# Podejście oparte na modelu

Survey Methodology, June 2018  
Vol. 44, No. 1, pp. 117-144  
Statistics Canada, Catalogue No. 12-001-X

117

## Model-assisted calibration of non-probability sample survey data using adaptive LASSO

Jack Kuang Tsung Chen, Richard L. Valliant and Michael R. Elliott<sup>1</sup>

### Abstract

The probability-sampling-based framework has dominated survey research because it provides precise mathematical tools to assess sampling variability. However increasing costs and declining response rates are expanding the use of non-probability samples, particularly in general population settings, where samples of individuals pulled from web surveys are becoming increasingly cheap and easy to access. But non-probability samples are at risk for selection bias due to differential access, degrees of interest, and other factors. Calibration to known statistical totals in the population provide a means of potentially diminishing the effect of selection bias in non-probability samples. Here we show that model calibration using adaptive LASSO can yield a consistent estimator of a population total as long as a subset of the true predictors is included in the prediction model, thus allowing large numbers of possible covariates to be included without risk of overfitting. We show that the model calibration using adaptive LASSO provides improved estimation with respect to mean square error relative to standard competitors such as generalized regression (GREG) estimators when a large number of covariates are required to determine the true model, with effectively no loss in efficiency over GREG when smaller models will suffice. We also derive closed form variance estimators of population totals, and compare their behavior with bootstrap estimators. We conclude with a real world example using data from the National Health Interview Survey.

**Key Words:** Adaptive LASSO estimators; Generalized regression estimator; Non-representative sample; Over-fitting; Variable selection; Oracle property.

# Podejście oparte na modelu



Journal of the Royal Statistical Society  
Applied Statistics  
Series C

*Appl. Statist.* (2019)  
**68**, Part 3, pp. 657–681

## Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling

Jack Kuang Tsung Chen

*SurveyMonkey, Palo Alto, USA*

and Richard L. Valliant and Michael R. Elliott

*University of Michigan, Ann Arbor, USA*

[Received December 2017. Final revision October 2018]

**Summary.** Declining response rates and increasing costs have led to greater use of non-probability samples in election polling. But non-probability samples may suffer from selection bias due to differential access, degrees of interest and other factors. Here we estimate voting preference for 19 elections in the US 2014 midterm elections by using large non-probability surveys obtained from SurveyMonkey users, calibrated to estimated control totals using model-assisted calibration combined with adaptive LASSO regression, or the estimated controlled LASSO, ECLASSO. Comparing the bias and root-mean-square error of ECLASSO with traditional calibration methods shows that ECLASSO can be a powerful method for adjusting non-probability surveys even when only a small sample is available from a probability survey. The methodology proposed has potentially broad application across social science and health research, as response rates for probability samples decline and access to non-probability sam-

# Podejście oparte na modelu

## Combining Non-probability and Probability Survey Samples Through Mass Imputation

Jae Kwang Kim<sup>1</sup>, Seho Park<sup>2</sup>, Yilin Chen<sup>3</sup>, and Changbao Wu<sup>3</sup>

<sup>1</sup> Department of Statistics, Iowa State University,

<sup>2</sup> Department of Biostatistics, Indiana University School of Medicine,

<sup>3</sup> Department of Statistics and Actuarial Science, University of Waterloo

**Summary.** Analysis of non-probability survey samples requires auxiliary information at the population level. Such information may also be obtained from an existing probability survey sample from the same finite population. Mass imputation has been used in practice for combining non-probability and probability survey samples and making inferences on the parameters of interest using the information collected only in the non-probability sample for the study variables. Under the assumption that the conditional mean function from the non-probability sample can be transported to the probability sample, we establish the consistency of the mass imputation estimator and derive its asymptotic variance formula. Variance estimators are developed using either linearization or bootstrap. Finite sample performances of the mass imputation estimator are investigated through simulation studies. We also address important practical issues of the method through the analysis of a real world non-probability survey sample collected by the Pew Research Centre.

**Keywords:** Auxiliary variables; Bootstrap variance estimator; Data integration; Ignorable sample selection; Model transportability; Selection bias

## Metody oparte na modelu

# Podejście oparte na modelu – klasyfikacja metod

- Gdy znamy wszystkie jednostki z populacji,
- Gdy znamy tylko jednostki z próby nielosowej – masowa imputacja (metoda Riversa, metody opracowane przez Jae-Kwang Kim'a i współpracowników).

# Podejście oparte na modelu – założenia

- Model z próby losowej możemy przenieść na resztę jednostek (ang. missing at random) – tzw. model dla super-populacji.
- Dysponujemy zmiennymi  $\mathbf{X}$ , które są obserwowane w próbie/próbach i populacji.
- Możemy zidentyfikować jednostki między próbą nielosową i populacją.
- Zakładamy, że zmienna  $Y$  oraz  $\mathbf{X}$  są obserwowane bez błędów.
- Zakładamy brak korelacji między obserwacjami, brak błędu nadreprezentacji itp.

## Metody oparte na modelu

## Masowa imputacja – dwa podejścia

	X1 (płeć)	X1 (wiek)	Y (słucha podcastów)	w (waga)	R	Y* (przepisane z nielosowej)	
Próba losowa	M	34	?	4	0	Nie	Ten zbiór wykorzystamy do estymacji
	M	32	?	2	0	Tak	
	K	50	?	5	0	Tak	
	K	40	?	10	0	Tak	
Próba nielosowa	M	32	Tak	?	1		
	M	34	Nie	?	1		
	K	40	Tak	?	1		
	K	50	Tak	?	1		

Rysunek 17: Podejście I: Przepisanie wartości z próby nielosowej

	X1 (płeć)	X1 (wiek)	Y (słuchanie podcastów)	w (waga)	R	$\hat{y}$ (przewidywana)	
Próba losowa	M	34	?	4	0	Nie	Ten zbiór wykorzystamy do estymacji
	M	32	?	2	0	Tak	
	K	50	?	5	0	Tak	
	K	40	?	10	0	Nie	
Próba nielosowa	M	32	Tak	?	1		
	M	34	Nie	?	1		
	K	40	Tak	?	1		
	K	50	Tak	?	1		

Rysunek 18: Podejście II: wartości przewidywane z modelu zbudowanego na próbie nielosowej

# Masowa imputacja – podejście I

W pierwszym podejściu, zaproponowanym przez Rivers (2007), dokonujemy masowej imputacji przez tzw. sample matching, który polega na następujących krokach:

- ➊ Dla próby nielosowej  $\mathcal{S}_A$  oraz losowej  $\mathcal{S}_B$  określamy zestaw wspólnych cech  $\mathbf{X}$ .
- ➋ Następnie, dla każdej jednostki  $i \in \mathcal{S}_B$  szukamy najbliższej jednostki ze zbioru  $k \in \mathcal{S}_A$ , tak żeby

$$d(\mathbf{x}_k, \mathbf{x}_i) = \|\mathbf{x}_k - \mathbf{x}_i\| \quad (28)$$

była jak najmniejsza. Możemy w tym celu wykorzystać np. odległość euklidesową. Jednostce  $i$  przypisujemy wartość cechy  $y_k$  ze zbioru  $\mathcal{S}_A$ .

- ➌ Po znalezieniu odpowiednich sąsiadów, wyznaczamy estymator. Przykładowo, estymator wartości średniej  $\theta = N^{-1} \sum_{i \in U} y_i$  dany będzie:

$$\hat{\theta}_{M1} = \sum_{i \in \mathcal{S}_B} w_i y_i / \sum_{i \in \mathcal{S}_B} w_i. \quad (29)$$

## Masowa imputacja – podejście I – UWAGA

Na podstawie pracy: Abadie & Imbens (2006) **Large sample properties of matching estimators for average treatment effects** (*Econometrica*, 74(1), 235-267) można wykazać, że obciążenie estymatora  $\hat{\theta}_{M1}$  rośnie wraz z liczbą zmiennych użytych do obliczenia  $d(\mathbf{x}_k, \mathbf{x}_i)$ . Dokładnie, ta zależność wyrażona jest następującym wzorem:

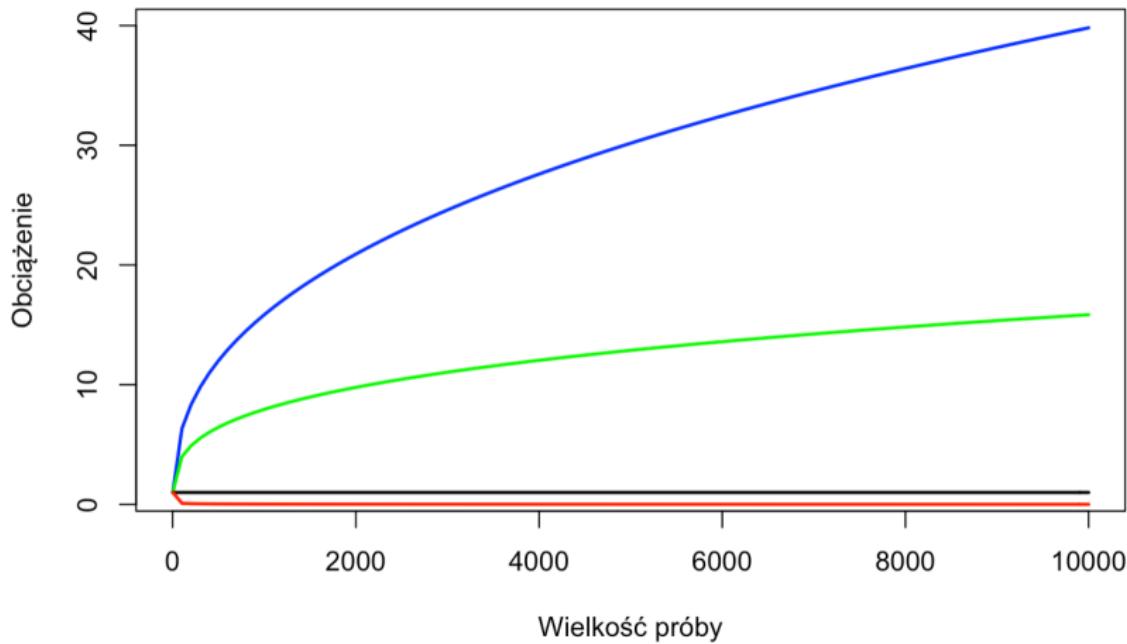
$$\text{Bias}(\hat{\theta}_{M1}) = O_p\left(n^{1/2 - 1/p}\right) \quad (30)$$

gdzie  $p$  oznacza liczbę zmiennych (wymiar wektora  $\mathbf{x}_k$ ), a  $O_p$  oznacza notację wielkie-O i określa asymptotyczne tempo wzrostu (od liczby próby losowej  $n_B$ ).

**Uwaga 1:** Oznacza to, że estymator  $\hat{\theta}_{M1}$  będzie nieobciążony wyłącznie gdy  $p = 1$ ,  $O_p(n^{-1/p}) = O_p(n^{-1/1}) = O_p(n^{-1})$  czyli obciążenie będzie mało wraz ze wzrostem próby losowej  $B$ .

**Uwaga 2:** ta zależność dotyczy zarówno prób nielosowych, jak i imputacji czy ekonometrycznego badania wpływu.

# Masowa imputacja – podejście I – obciążenie



Rysunek 19: Wizualizacja obciążenia  $O_p(n^{1/2-1/p})$ . Kolor czerwony:  $p=1$ ; czarny:  $p=2$ , zielony:  $p=5$  i niebieski:  $p=10$ .

# Masowa imputacja – podejście I – Praktyka

Mając na uwadze fakt, że przypisania wartości  $y_i$  z próby nielosowej dla jednostek  $k$  z próby losowej należy wykorzystać na podstawie wyłącznie jednej zmiennej, stosuje się następujące podejście:

- ➊ Budujemy model  $m(\mathbf{x}_i; \boldsymbol{\beta})$  na próbie nielosowej  $\mathcal{S}_A$  otrzymując  $\hat{y}_i = m(\mathbf{x}_i; \hat{\boldsymbol{\beta}})$ .
- ➋ Stosujemy model  $m(\mathbf{x}; \hat{\boldsymbol{\beta}})$  na próbie losowej  $\mathcal{S}_B$  otrzymując  $\hat{y}_k = m(\mathbf{x}_k; \hat{\boldsymbol{\beta}})$ .
- ➌ Dla każdej jednostki  $k \in \mathcal{S}_B$  znajdujemy najbliższą jednostkę na podstawie  $d(\hat{y}_k, \hat{y}_i)$  i przepisujemy wartość  $y_i$ .
- ➍ Następnie wyznaczamy estymator

$$\hat{\theta}_{M1} = \sum_{i \in \mathcal{S}_B} w_i y_i / \sum_{i \in \mathcal{S}_B} w_i. \quad (31)$$

To podejście w literaturze nazywa się *predictive mean matching*.

## Masowa imputacja – podejście II

- W pracy Kim, Park, Chen i Wu (2021) **Combining Non-probability and Probability Survey Samples Through Mass Imputation**, (Journal of the Royal Statistical Society: Series A) zaproponowano trochę inne podejście ale oparte na solidnych, teoretycznych podstawach.
- Zamiast dokonywać poszukiwania najbliższego sąsiada wykorzystuje się wyłącznie 1 oraz 2 krok z poprzedniego slajdu.
- Estymator wartości średniej dany jest wtedy

$$\hat{\theta}_{M2} = \sum_{i \in S_B} w_i \hat{y}_i / \sum_{i \in S_B} w_i. \quad (32)$$

- Ten sposób nazywamy na potrzeby zajęć podejściem II do masowej imputacji.
- W wyżej wymienionej pracy zaproponowano również estymator wariancji w postaci zlinearyzowanej (konkretny wzór), jak i na podstawie metody bootstrap.

# Masowa imputacja – przykład empiryczny

Przejdźmy do przykładu

# Podwójnie odporne estymatory

- Propensity score weighting działa **wyłącznie wtedy, gdy** model  $P(R_i = 1|\mathbf{x}_i)$  jest poprawnie **Wyspecyfikowany** – zakładany model jest poprawny dla całej populacji.
- Dlatego w literaturze zaproponowano nową klasę estymatorów pod nazwą: podwójnie odporne estymatory (ang. *double robust estimators*).
- Dlaczego podwójnie odporne? Estymator ten składa się z dwóch części

$$\hat{\theta}_{\text{DR}} = \underbrace{\frac{1}{N} \sum_{i=1}^N \frac{R_i \left\{ y_i - m(\hat{\beta}; \mathbf{x}_i) \right\}}{\rho(\hat{\lambda}; \mathbf{x}_i)}}_{\text{Średnia ważona reszt z modelu}} + \underbrace{\frac{1}{N} \sum_{i=1}^N m(\hat{\beta}; \mathbf{x}_i)}_{\text{Średnia z predykcji dla całej populacji}}, \quad (33)$$

gdzie  $R_i$  to zmienna 0-1, gdzie 1 gdy próba nielosowa,  $\rho(\hat{\lambda}; \mathbf{x}_i)$  to prawdopodobieństwo przynależności do próby nielosowej, a  $m(\hat{\beta}; \mathbf{x}_i)$  to pewien model parametryczny ( $E(y|\mathbf{x}) = m(\hat{\beta}; \mathbf{x}_i)$ )).

[Podwójnie odporne estymatory](#)

# Podwójnie odporne estymatory – w telegraficznym skrócie

**Do wartości przewidywanych dla jednostek spoza próby nielosowej dodajemy ważone reszty z modelu dla próby nielosowej.**

## Podwójnie odporne estymatory

# Podwójnie odporne estymatory

Właściwości:

- Estymator ten jest nieobciążony gdy model dla  $\rho(\hat{\lambda}; \mathbf{x}_i)$  jest źle wyspecyfikowany, ale  $m(\hat{\beta}; \mathbf{x}_i)$  jest dobrze wyspecyfikowany,
- Estymator ten jest nieobciążony gdy  $m(\hat{\beta}; \mathbf{x}_i)$  jest źle wyspecyfikowany ale  $\rho(\hat{\lambda}; \mathbf{x}_i)$  jest dobrze wyspecyfikowany.

Literatura:

- Chen, Y., Li, P., & Wu, C. (2020). Doubly Robust Inference With Nonprobability Survey Samples. *Journal of the American Statistical Association*, 115(532), 2011–2021.
- Kim, J. K., & Wang, Z. (2019). Sampling Techniques for Big Data Analysis. *International Statistical Review*, 87(S1), S177–S191.

## Podwójnie odporne estymatory

# Podwójnie odporne estymatory

Powyższe wzory miały zastosowanie gdy znamy wszystkie jednostki z populacji. Jednak gdy dysponujemy wyłącznie dwiema próbami (losową i nie-losową) estymator ten może mieć postać:

$$\hat{\theta}_{\text{DR1}} = \underbrace{\frac{1}{N} \sum_{i \in S_A} w_i^* \left\{ y_i - m(\mathbf{x}_i, \hat{\beta}) \right\}}_{\text{Próba nielosowa}} + \underbrace{\frac{1}{N} \sum_{i \in S_B} d_i m(\mathbf{x}_i, \hat{\beta})}_{\text{Próba losowa}}, \quad (34)$$

gdzie  $N$  to znana wielkość populacji,  $S_A$  to próba nielosowa,  $S_B$  to próba losowa,  $w_i^* = 1/\rho(\hat{\lambda}; \mathbf{x}_i)$ ,  $d_i$  to waga wynikająca z losowania.

## Podwójnie odporne estymatory

# Podwójnie odporne estymatory

W przypadku gdy wielkość populacji nie jest znana możemy zastosować następujący estymator

$$\hat{\theta}_{\text{DR2}} = \underbrace{\frac{1}{\hat{N}^A} \sum_{i \in S_A} w_i^* \left\{ y_i - m(\mathbf{x}_i, \hat{\beta}) \right\}}_{\text{Próba nielosowa}} + \underbrace{\frac{1}{\hat{N}^B} \sum_{i \in S_B} d_i m(\mathbf{x}_i, \hat{\beta})}_{\text{Próba losowa}}, \quad (35)$$

gdzie  $\hat{N}^A = \sum_{i \in S_A} w_i^*$ ,  $\hat{N}^B = \sum_{i \in S_B} d_i$ .

# Podwójnie odporne estymatory – estymacja wariancji

- Kim & Wang (2019) pokazali, że jeżeli próba losowa stanowi niewielki ułamek próby nielosowej ( $n_B/N_A = o(1)$ ) to wariancję estymatora  $\hat{\theta}_{DR1}$  lub  $\hat{\theta}_{DR2}$  można wyznaczyć wyłącznie na podstawie próby losowej (zgodnie z jej schematem losowania).
- Chen, Li & Wu (2020) wyznaczyli estymatory wariancji bez takiego założenia oraz zaproponowali podejście oparte na metodzie bootstrap, którą można scharakteryzować następującymi niezależnymi krokami:
  - ❶ dla próby nielosowej  $S_A$  losujemy ze zwracaniem prostą próbę  $S_A^*$  o liczebności  $n_A$ ,
  - ❷ dla próby losowej  $S_B$  losujemy z prawdopodobieństwem  $1/d_i^B$  ze zwracaniem próbę  $S_B$  o liczebności  $n_B$

Następnie wyznaczamy  $\hat{\theta}_{DR1}^*$  lub  $\hat{\theta}_{DR2}^*$  z każdej próby bootstrapowej.

# Podwójnie odporne estymatory – rozszerzenie

- W powyższych pracach nie rozważano kwestii doboru zmiennych do obydwu modeli,
- Praca: *Yang, S., Kim, J. K., & Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82(2), 445-465* przedstawia rozwiązanie w tym zakresie oparte na doborze zbliżonym do regresji LASSO (dokładnie Smoothly Clipped Absolute Deviation; SCAD). Jest również pakiet w R ale o dość ograniczonych możliwościach.

**Podwójnie odporne estymatory**

# Podwójnie odporne estymatory – przykład empiryczny

Przejdźmy do przykładu