

Estymacja na podstawie prób nielosowych

dr Maciej Beręsewicz, prof. UEP

Katedra Statystyki, Uniwersytet Ekonomiczny w Poznaniu

Ośrodek Statystyki Małych Obszarów, Urząd Statystyczny w Poznaniu

Spis treści

1 Wprowadzenie

2 Internet w Polsce

3 Reprezentatywność

- Reprezentatywność – definicje, pomiar, problematyka

4 Metody estymacji dla prób nielosowych

- Metody quasi-randomizacyjne
 - Post-stratyfikacja
 - Kalibracja
 - Propensity score/inverse probability weighting
- Metody oparte na modelu
 - Ogólna idea
 - Podwójnie odporne estymatory

Spis treści

- 1 Wprowadzenie
- 2 Internet w Polsce
- 3 Reprezentatywność
- 4 Metody estymacji dla prób nielosowych

Oczekiwania

Oczekiwania

Proszę wejść na stronę www.menti.com,
wpisać kod: **4646 6185**,
i podać swoje oczekiwania względem szkolenia.

Test wiedzy

Test wiedzy

Proszę wejść na stronę www.menti.com,
wpisać kod: **2344 8310**,
i odpowiedzieć na pytania!

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 Wprowadzenie
- 2 Internet w Polsce
- 3 Reprezentatywność
- 4 Metody estymacji dla prób nielosowych

Internet w Polsce

<https://forms.gle/nF1aL9z2WCZAawp9A>

Internet w Polsce

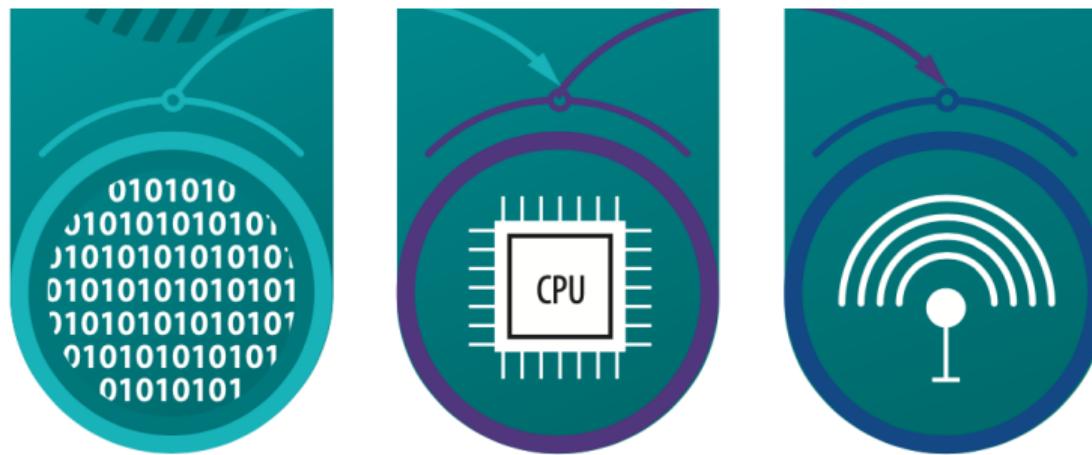
Krótki quiz wiedzy o Internecie w Polsce

Proszę wejść na stronę www.menti.com,
wpisać kod: **8662 6996**,
a następnie odpowiedzieć na dwa pytania.

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



Warszawa, Szczecin 2022

Społeczeństwo informacyjne w Polsce w 2022 r.

Information society in Poland in 2022

Internet w Polsce – źródło: badanie ICT GUS

Gospodarstwa domowe posiadające dostęp do Internetu w domu

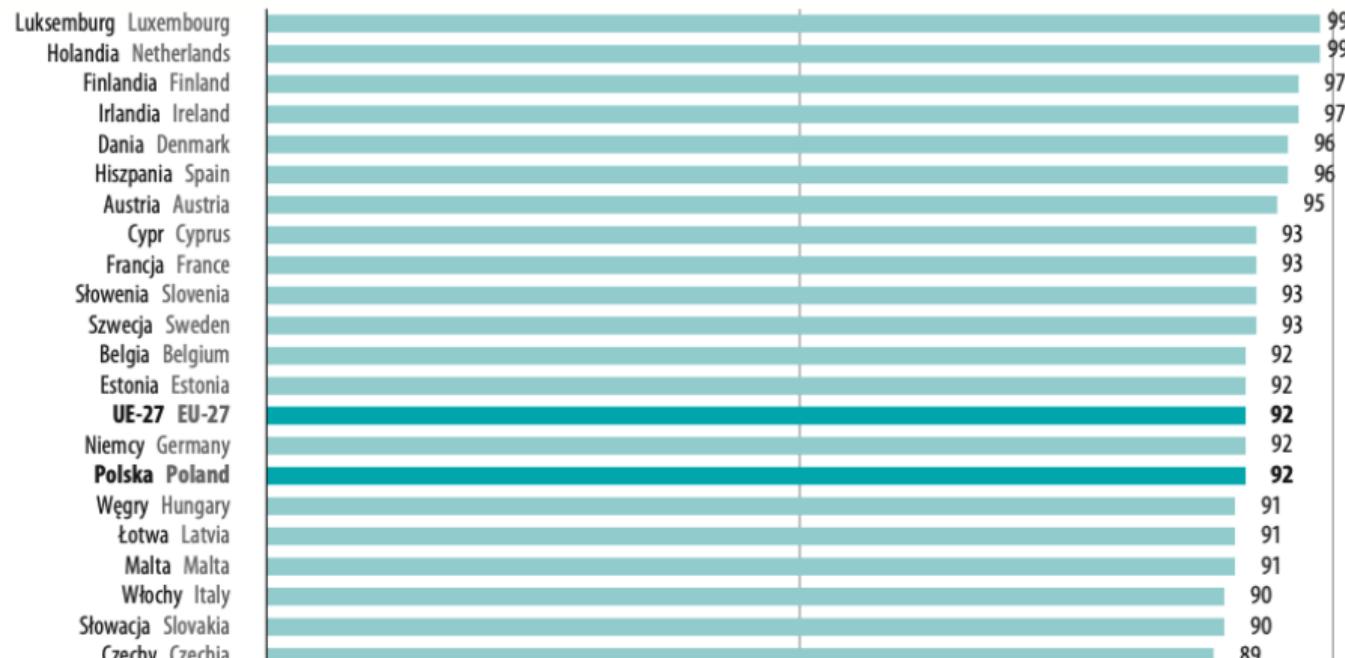
Households with access to the Internet at home

Wyszczególnienie Specification		2018	2019	2020	2021	2022
		w % ogółu gospodarstw danej grupy in % of total households in a group				
Ogółem	Total	84,2	86,7	90,4	92,4	93,3
Typ gospodarstwa domowego Household type						
Gospodarstwa z dziećmi Households with children		99,2	99,3	99,5	99,7	99,9
Gospodarstwa bez dzieci Households without children		77,0	80,4	85,9	88,8	90,5
Miejsce zamieszkania Domicile						
Duże miasta Large cities		87,8	90,0	92,1	93,8	94,4
Mniejsze miasta Small cities		82,7	85,6	89,7	91,6	92,3
Obszary wiejskie Rural areas		82,0	84,6	89,3	91,8	93,2
Stopień urbanizacji Degree of urbanisation						
Niski Thinly populated		81,6	83,5	88,9	91,9	92,8

Internet w Polsce – źródło: badanie ICT GUS

Gospodarstwa domowe z dostępem do Internetu w domu w krajach Unii Europejskiej w 2021 r.

Households with access to the Internet at home in European Union countries in 2021



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		2018	2019	2020	2021	2022
		w % ogółu gospodarstw danej grupy in % of total households in a group				
Ogółem	Total	79,3	83,3	89,6	91,7	92,6
Typ gospodarstwa domowego Household type						
Gospodarstwa z dziećmi Households with children		95,0	95,9	99,1	99,4	99,6
Gospodarstwa bez dzieci Households without children		71,8	77,0	84,9	87,9	89,6
Miejsce zamieszkania Domicile						
Duże miasta Large cities		83,4	87,1	91,0	93,1	93,4
Mniejsze miasta Small cities		78,2	81,9	89,1	91,2	91,6
Obszary wiejskie Rural areas		76,2	80,7	88,7	90,9	92,8
Stopień urbanizacji Degree of urbanisation						
Najniższa		75,7	79,0	88,0	91,1	92,5
Próby niesosowe (Beręsewicz)		75,7	79,0	88,0	91,1	92,5
Estymacja na podstawie prób niesosowych		75,7	79,0	88,0	91,1	92,5

Internet w Polsce – źródło: badanie ICT GUS

Osoby regularnie korzystające z Internetu w krajach Unii Europejskiej w 2021 r.
Regular Internet users in European Union countries in 2021



Internet w Polsce – źródło: badanie ICT GUS

Częstotliwość korzystania z Internetu

Frequency of Internet use

Wyszczególnienie Specification	2018	2019	2020	2021	2022
-----------------------------------	------	------	------	------	------

W % ogółu osób In % of total individuals

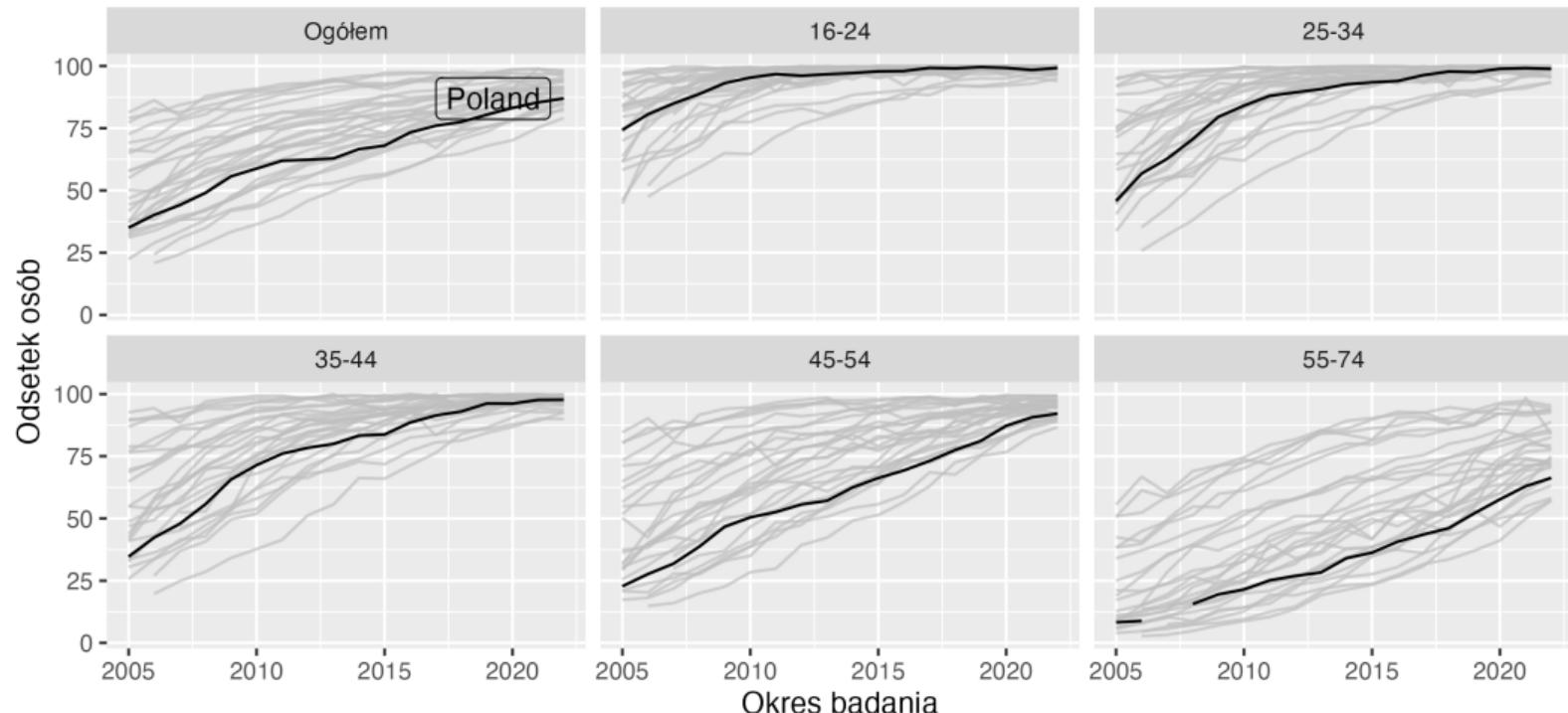
Regularnie Regularly	74,8	78,3	81,4	83,6	85,7
Codziennie lub prawie codziennie Every day or almost every day	63,9	68,2	72,3	73,7	80,3
Przynajmniej raz w tygodniu, ale nie każdego dnia At least once a week but not every day	10,9	10,1	9,0	10,0	5,4
Rzadziej niż raz w tygodniu Less than once a week	2,8	2,2	1,8	1,7	1,2

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

Regularnie Regularly	96,4	97,3	97,8	98,0	98,6
Codziennie lub prawie codziennie Every day or almost every day	82,4	84,8	87,0	86,3	92,4
Przynajmniej raz w tygodniu, ale nie każdego dnia At least once a week but not every day	14,0	12,6	10,9	11,7	6,2
Rzadziej niż raz w tygodniu Less than once a week	2,6	2,7	2,2	2,0	1,4

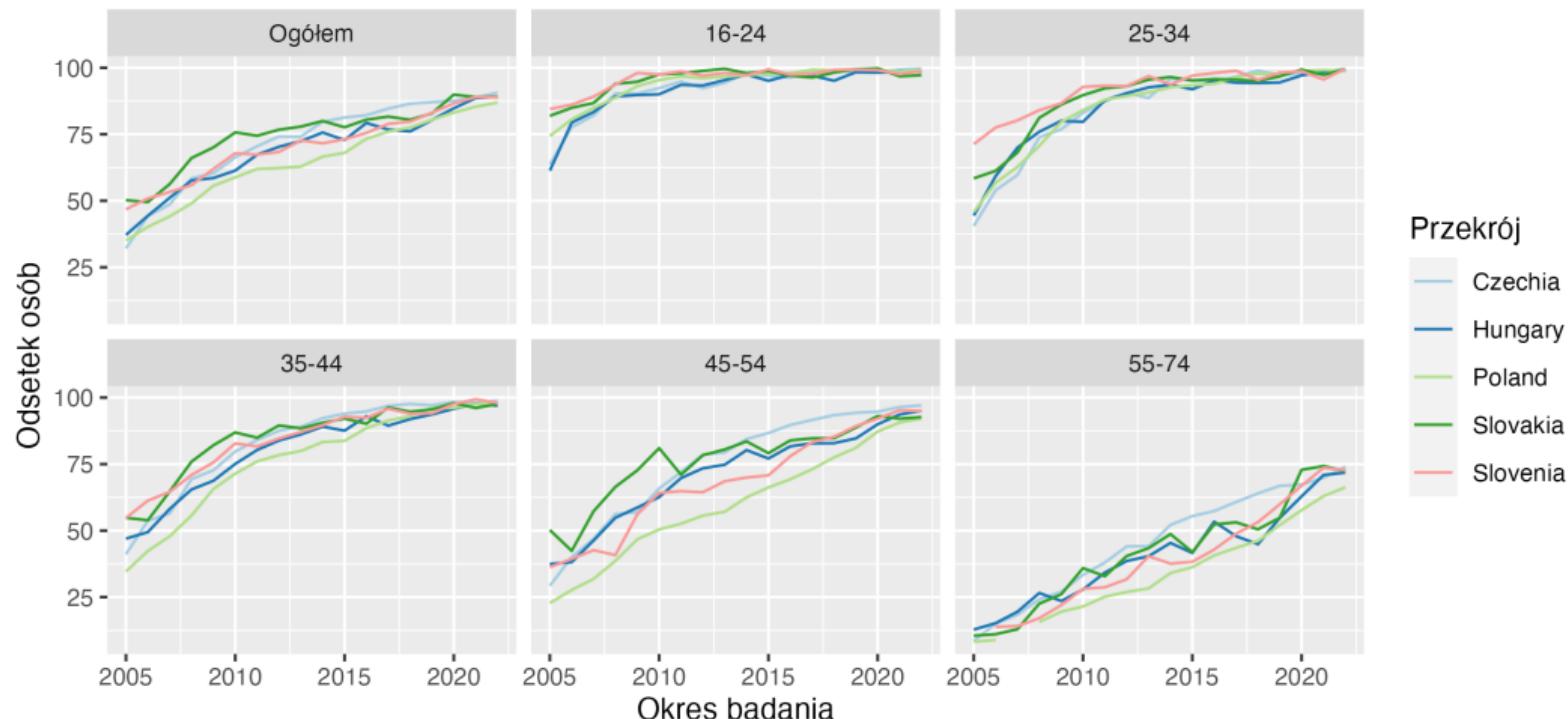
Internet w Polsce – źródło: badanie ICT GUS

Odsetek korzystających z Internetu w ostatnich 3 miesiącach w Polsce i UE



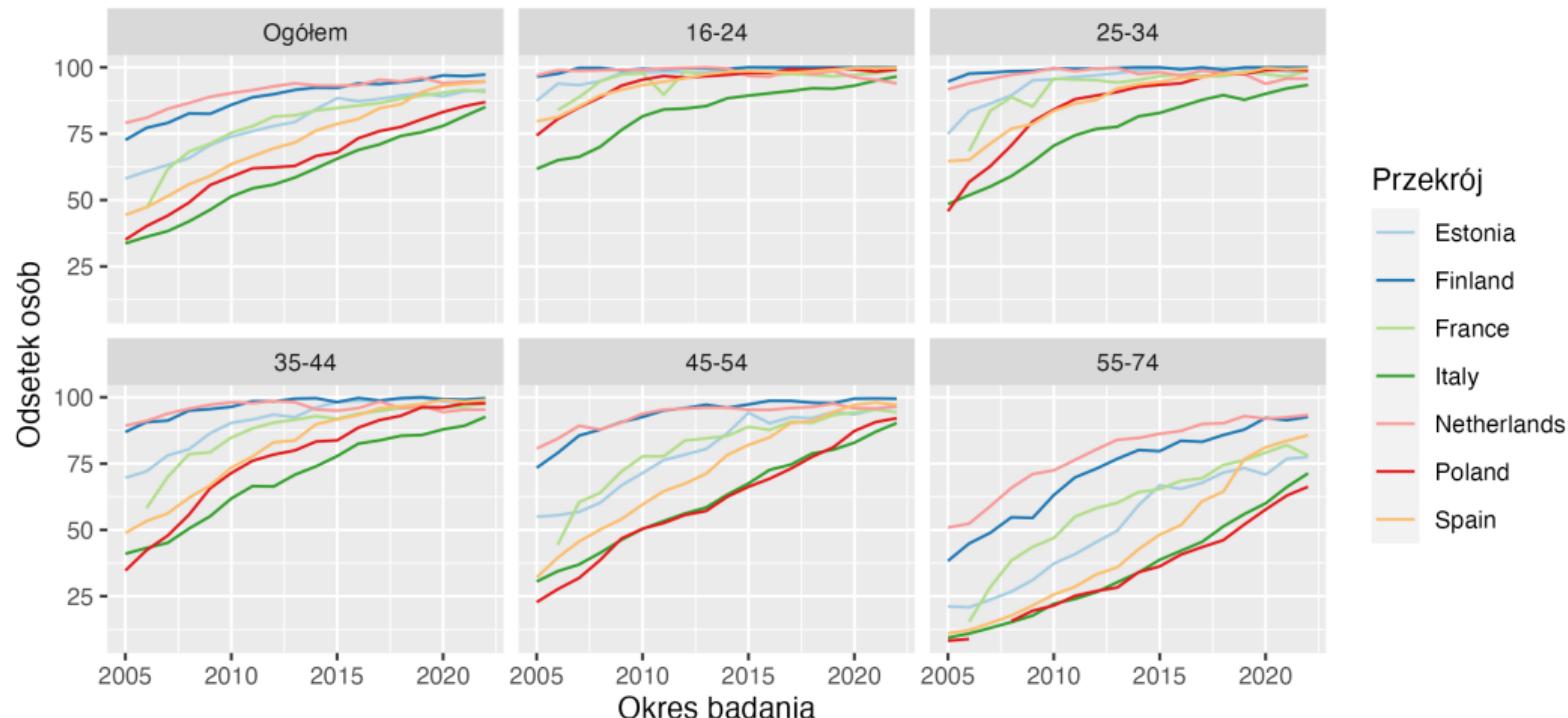
Internet w Polsce – źródło: badanie ICT GUS

Odsetek korzystających z Internetu w ostatnich 3 miesiącach w Polsce i wybrane kraje



Internet w Polsce – źródło: badanie ICT GUS

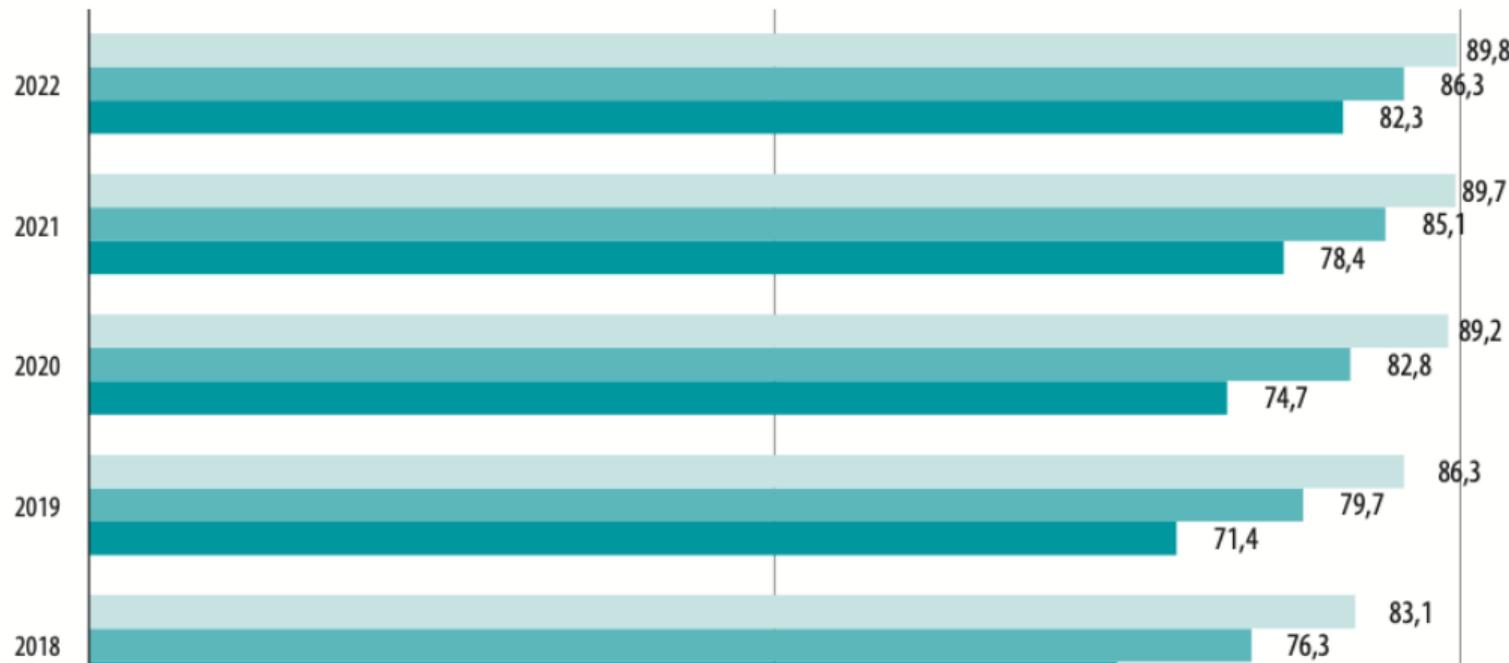
Odsetek korzystających z Internetu w ostatnich 3 miesiącach w Polsce i wybrane kraje



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		2018	2019	2020	2021	2022
		w % ogółu osób danej grupy in % of total individuals in a group				
16–24 lata	16–24 years	98,8	99,3	99,2	98,4	99,0
25–34		96,5	97,0	98,4	98,9	98,7
35–44		90,6	94,5	95,2	96,7	97,1
45–54		73,4	78,1	84,3	89,1	91,1
55–64		50,4	59,9	65,8	71,3	75,5
65–74 lata	65–74 years	29,8	33,3	40,4	45,9	51,0

Internet w Polsce – źródło: badanie ICT GUS

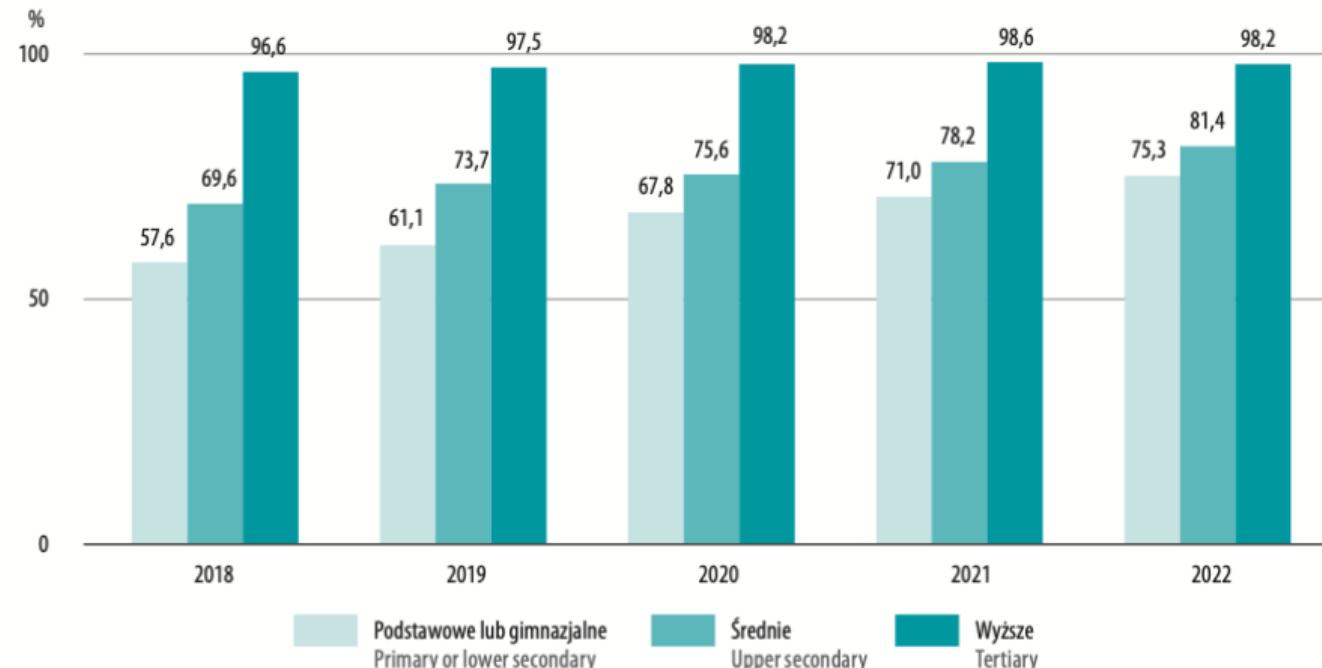
Gospodarstwa domowe posiadające dostęp do Internetu w domu oraz osoby korzystające z Internetu według województw w 2022 r.

Households with access to the Internet at home and Internet users by voivodships in 2022

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
Polska Poland	93,3	90,6	85,7
Dolnośląskie	93,5	91,9	88,6
Kujawsko-pomorskie	93,8	92,7	87,9
Lubelskie	92,4	87,8	81,3
Lubuskie	93,7	88,6	82,6
Łódzkie	91,3	90,3	86,6
Małopolskie	93,3	91,2	86,1

Internet w Polsce – źródło: badanie ICT GUS

Osoby regularnie korzystające z Internetu według poziomu wykształcenia Regular Internet users by educational level



Internet w Polsce – źródło: badanie ICT GUS

Osoby regularnie korzystające z Internetu według aktywności zawodowej

Regular Internet users by employment situation

Wyszczególnienie Specification	2018	2019	2020	2021	2022
	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	43,1	48,3	53,5	57,6	60,4
Bezrobotni Unemployed	65,7	72,0	82,6	83,2	86,3
Pracujący Persons employed	84,9	89,8	91,8	93,3	95,2
Rolnicy Farmers	57,8	60,0	65,1	75,1	83,0
Pracujący na własny rachunek Self-employed	92,6	95,5	96,0	98,0	96,7
Pracownicy najemni Employees	88,7	92,0	93,6	94,8	96,1
Uczniowie i studenci Students	99,6	99,6	99,8	99,1	99,3

Internet w Polsce – źródło: badanie ICT GUS

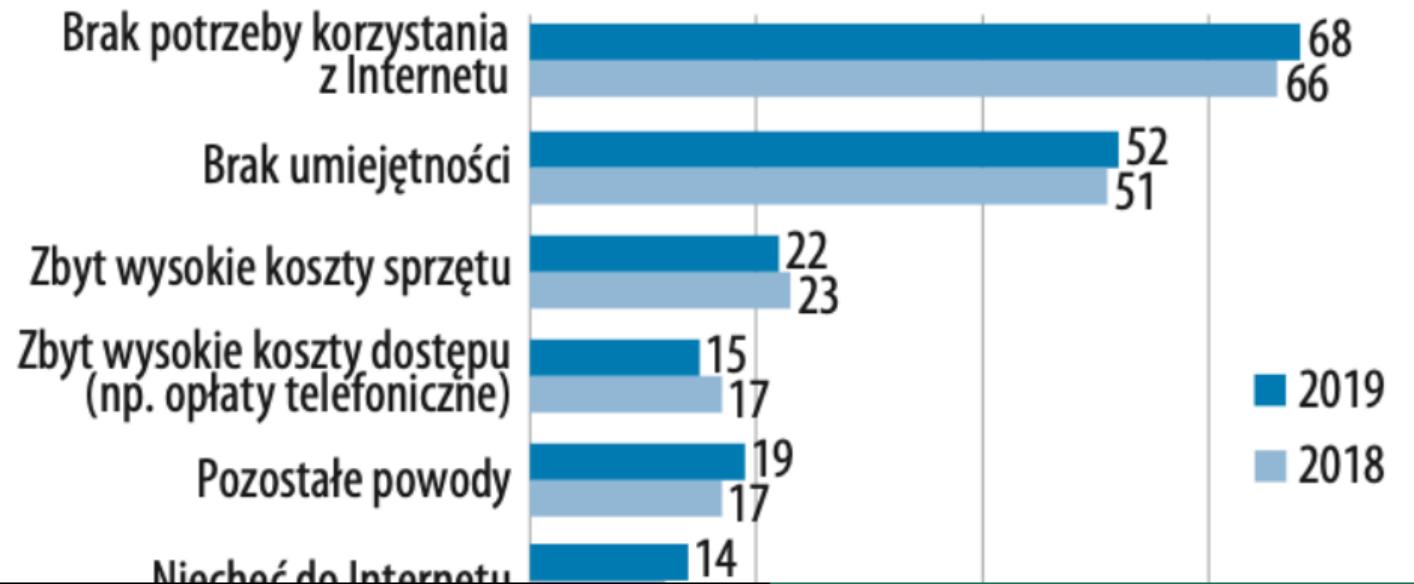
Osoby korzystające z Internetu w sprawach prywatnych w ciągu ostatnich 3 miesięcy według wybranych celów

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

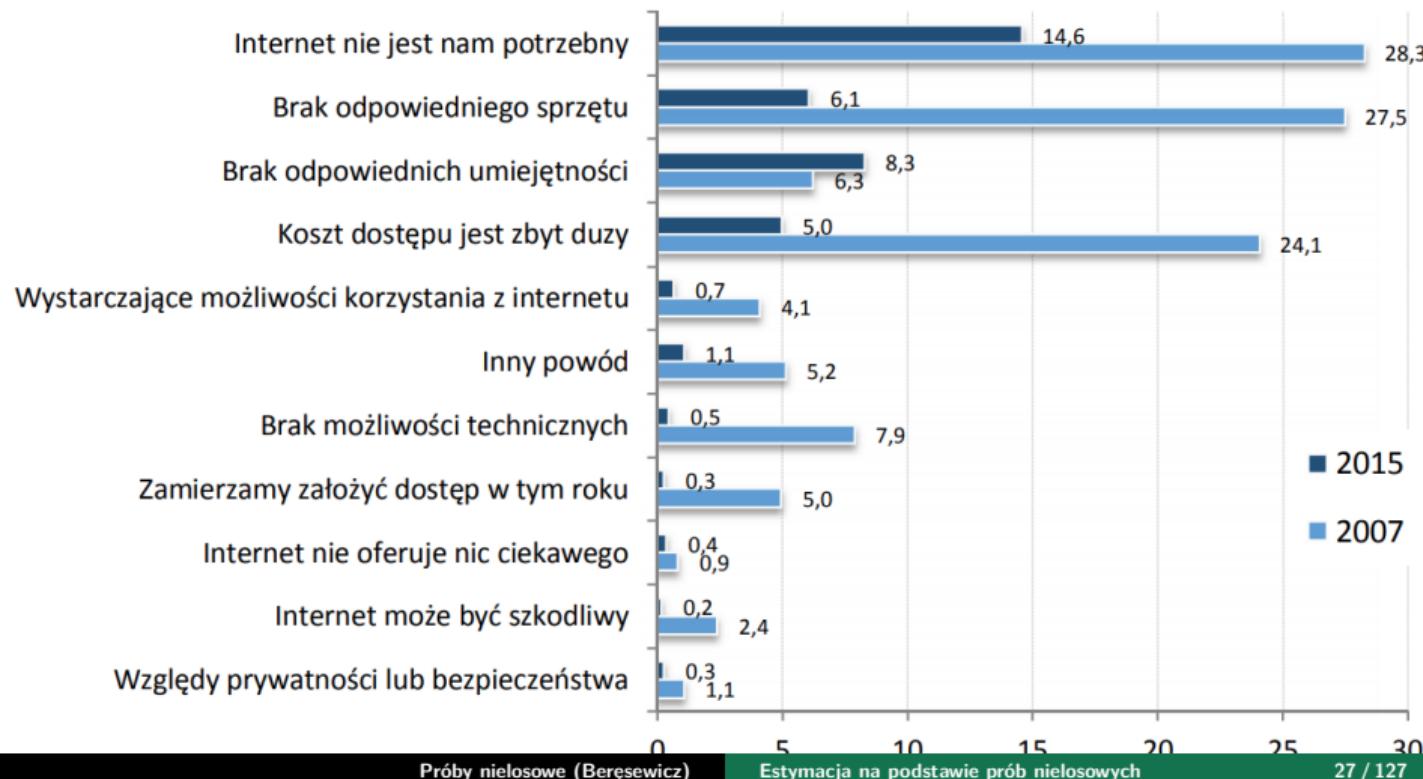
Cele korzystania z Internetu Purposes of Internet usage	2018	2019	2020	2021	2022	2018	2019	2020	2021	2022
	w % ogółu osób in % of total individuals					w % osób korzystających z Internetu in % of Internet users				
Korzystanie z poczty elektronicznej Sending, receiving e-mail	60,7	64,8	65,9	68,3	69,3	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 bankowczych	44,0	47,3	49,5	52,2	55,6	56,8	58,8	59,5	61,2	63,9

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.)

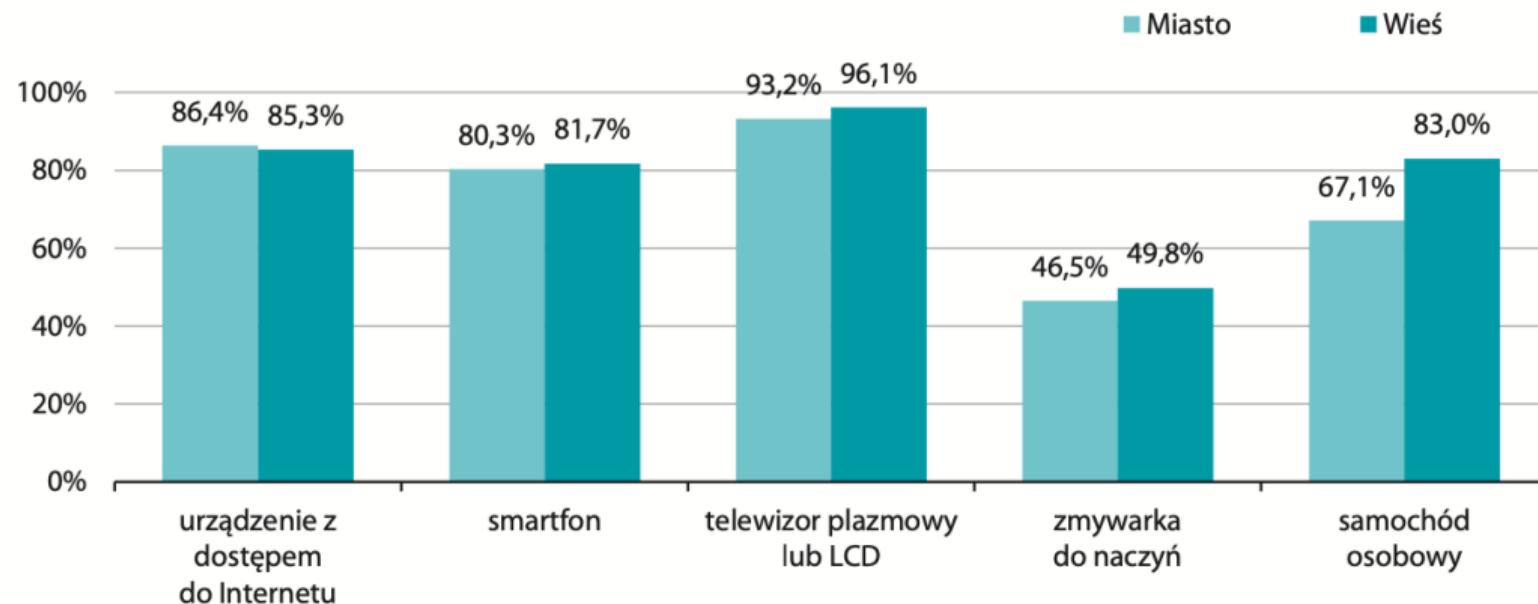


Internet w Polsce – źródło: Diagnoza Społeczna 2015



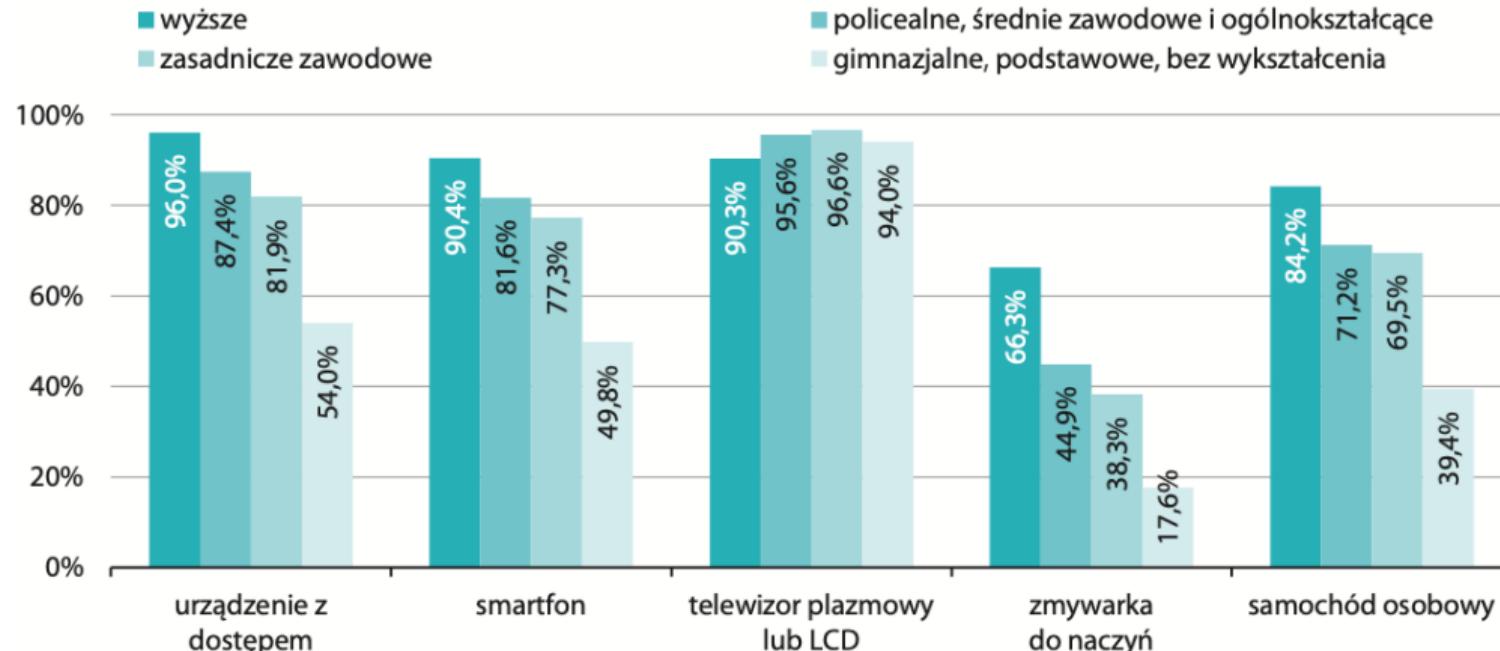
Internet w Polsce – źródło: BBGd 2022

Wyposażenie gospodarstw domowych w wybrane dobra trwałego użytkowania według miejsca zamieszkania w 2021 r.



Internet w Polsce – źródło: BBGd 2022

Wyposażenie gospodarstw domowych w wybrane dobra trwałego użytkowania według poziomu wykształcenia osoby odniesienia w 2021 r.



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

Spis treści

1 Wprowadzenie

2 Internet w Polsce

3 Reprezentatywność

- Reprezentatywność – definicje, pomiar, problematyka

4 Metody estymacji dla prób nielosowych

Zadanie – 5 minut

Której wersji bardziej Państwo zaufacie? Dlaczego?

- ① 1% próbie losowej z 60% realizacją (odsetkiem odpowiedzi), czy
- ② 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 θ , to

- estymator jest nieobciążony gdy

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

- obciążenie estymatora oznaczamy przez

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

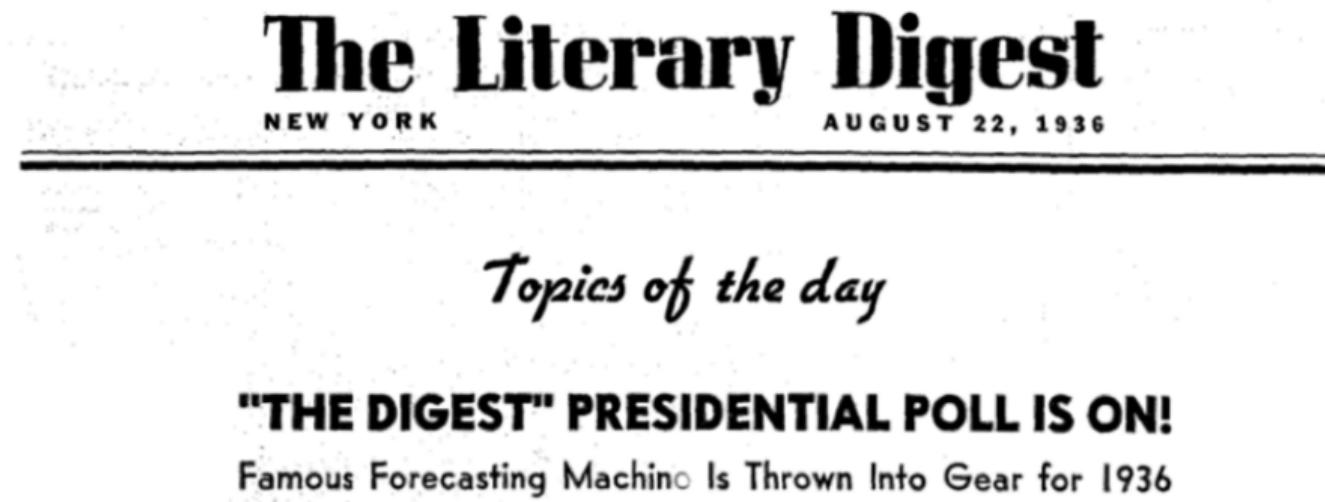
- wariancję estymatora natomiast

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

- 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 tygodniakiem.
- 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

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 (zmienne) 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-stosowane-w-statystyce/2771_pojecie.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-stosowane-w-statystyce/2748_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-stosowane-w-statystyce-nielosowej.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, rejesty statystyczne, sprawozdawczość przedsiębiorstw)
- ② **Źródła niestatystyczne** – wszystkie pozostałe źródła danych, których celem nie jest dostarczanie informacji statystycznej o całej populacji (np. rejesty 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. Citro¹

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 best being 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, consumer and business expenditures, et al. Faced with secularly declining unit and item response rates and evidence of reporting error, we have responded in many ways, 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 user needs from multiple data sources. Such sources include administrative records and, increasingly, transaction and Internet-based data. I provide two examples - household income and plumbing 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

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

Przykład – CBOP



Język: PL



Kontrast:



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 praktykiLiczba 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

PRACODAWCA

DOSTĘPNA OD

[SPRZEDAWCA](#)Bytom,
śląskie

Umowa o pracę

kontakt przez PUP

dzisiaj

WYBRANE KRYTERIUM

Przykład – OtoDom

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

Mieszkania ▾ na sprzedaż ▾ **Cały kraj: Polska** [Wyszukaj](#)

Cena ▾ Powierzchnia ▾ Obsługa zdalna NOWE ▾ Liczba pokoi ▾ Rynek ▾ Więcej filtrów ▾

Mieszkania na sprzedaż [Zapisz wyszukiwanie](#) [Lista](#) [Mapa](#)

liczba ofert: **140 350** sortuj po: domyślnie ▾ [Zobacz wszystkie](#)

Promowane ogłoszenia



3 Pokoje / Osiedle Leśne / M. W Hali / Premium
Mieszkanie na sprzedaż: Bydgoszcz, Osiedle Leśne
[Dodaj do ulubionych](#)

3 pokoje 77,67 m² 8 279 zł/m² **643 000 zł**

Nasalski Nieruchomości i Finanse

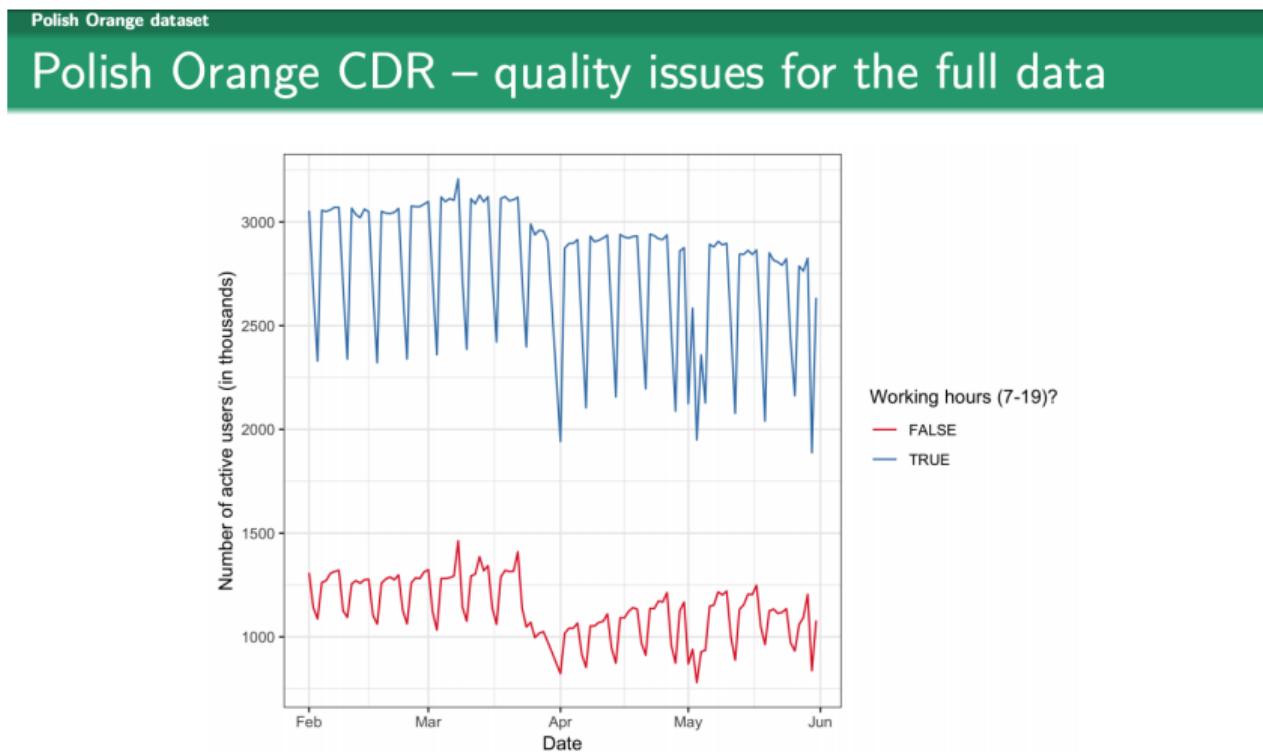
Zobacz też oferty deweloperów



od 6 000 zł/m²
Szczecin, Zachód, Krzekowo **DOM**
Zamknięte osiedle domków na Krzekowie
Nowe-domy.com Przedstawiciel dewelopera

[Próby niesosowe \(Beręsewicz\)](#) [Estymacja na podstawie prób niesosowych](#) 50 / 127

Przykład – Kałużny, Beręsewicz i Filipowska (2018)



Przejdźmy do precyzyjnych definicji reprezentatywności

Podstawowe oznaczenia

- $I_i = \{0, 1\}$ – zmienna indykatorkowa; określa czy dana jednostka i była obserwowana w określonym zbiorze (np. ma dostęp do Internetu; ang. *inclusion*).
- $R_i = \{0, 1\}$ – zmienna indykatorkowa; 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 ; 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

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

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

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

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

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

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 ρ , a Y , σ 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 ELECTION¹

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

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 ρ 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. $f = n/N$,
- σ_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 ρ 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

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

Wybory w USA – Clinton vs Trump

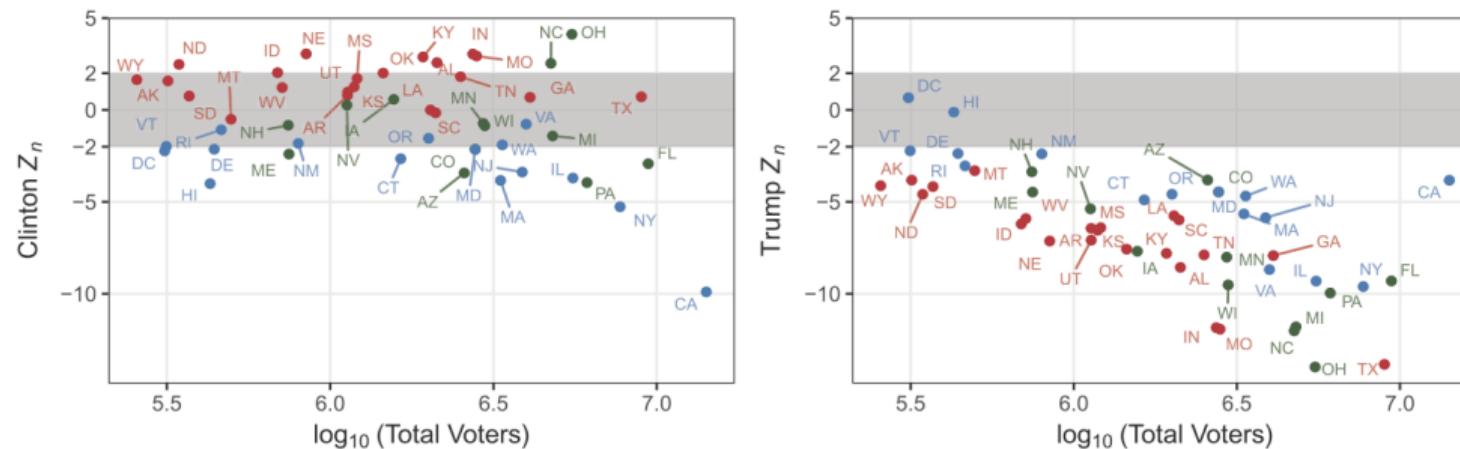


FIG. 7. Estimates of $Z_n = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}}$: The conventional 95% confidence interval region $|Z_n| \leq 2$ is indicated in gray.

Rysunek 3: 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 4: 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零u 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 Wprowadzenie

2 Internet w Polsce

3 Reprezentatywność

4 Metody estymacji dla prób nielosowych

- Metody quasi-randomizacyjne
- Metody oparte na modelu

Metody estymacji – rozważmy następujące przypadki

A)

	X	Y
Próba losowa		
Próba nieosowa		

C)

	X	Y*	Y
Próba losowa			
Próba nieosowa			

B)

	X	Y
Próba losowa		
Próba nieosowa		

D)

	X	Y
Rejestr / spis jednostek		
Próba nieosowa		

Rysunek 5: 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.

Elliott i Valliant (2017) wyróżniają dwa podejścia:

- **quasi-randomizacyjne** – w której konstruujemy *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				

Metody quasi-randomizacyjne

W przypadku metod quasi-randomizacyjnych możemy rozważyć następujące metody:

- **Post-stratyfikację** (ang. post-stratification) – wymaga znajomości wartości globalnych X (Holt & Smith, 1979)
- **Kalibrację** (ang. calibration) – wymaga znajomości wartości globalnych / średnich X (Deville & Särndal, 1992)
- **Ważenie przez dopasowanie** (ang. propensity score weighting) – wymaga dostępu do danych jednostkowych lub wartości globalnych cech X (Lee, 2006)

Literatura:

- Holt, D., & Smith, T. F. (1979). Post stratification. *Journal of the Royal Statistical Society: Series A (General)*, 142(1), 33-46.
- Deville, J. C., & Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418), 376-382.
- Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of official statistics*, 22(2), 329.

Post-stratyfikacja

Podstawowe oznaczenia

- Niech zmienna X ma H poziomów (np. zmienna województwo ma 16 poziomów).
- Zmienna X dzieli populację Ω na H warstw tj. $\Omega_1, \Omega_2, \dots, \Omega_H$.
- Liczba jednostek populacji Ω_H wynosi N_H , czyli $N = N_1 + N_2 + \dots + N_H$,
- Zakładamy, że zbiór big data (s) ma wielkość n i też można go podzielić według zmiennej X na H poziomów. To jest $n = n_1 + n_2 + \dots + n_H$.
- Następnie dla każdej jednostki i ze zbioru próby nielosowej (s) przypisujemy wagę utworzoną w następujący sposób

$$w_i = \frac{N_h/N}{n_h/n} \quad (20)$$

- Następnie wagę w_i wykorzystujemy do wyznaczenia średniej ważonej w warstwie.

Post-stratyfikacja – estymator

Post-stratyfikacyjny estymator średniej cechy Y dla próby big data s jest opisany następującym wzorem

$$\bar{y}_{s,PS} = \frac{1}{n} \sum_{i=1}^N s_i w_i l_i y_i = \frac{1}{N} \sum_{h=1}^H N_h \bar{y}_s^{(h)} = \sum_{h=1}^H W_h \bar{y}_s^{(h)}, \quad (21)$$

gdzie $W_h = N_h/N$, a $\bar{y}_s^{(h)}$ to średnia arytmetyczna w warstwie.

Post-stratyfikacja – własności

- Wariancja tego estymatora dana jest wzorem

$$V(\bar{y}_{PS}) = \left(\frac{1}{n} - \frac{1}{N} \right) \sum_{h=1}^L W_h S_h^2 + \frac{1}{n^2} \sum_{h=1}^L (1 - W_h) S_h^2,$$

gdzie S_h^2 to wariancja y w warstwie h .

- Obciążenie tego estymatora w przypadku błędu auto-selekcji (ang. self-selection) można opisać następującym wzorem

$$B(\bar{y}_{S,PS}) = \sum_{h=1}^L W_h \frac{R_{\rho Y}^{(h)} S_{\rho}^{(h)} S_Y^{(h)}}{\bar{\rho}^{(h)}} \quad (22)$$

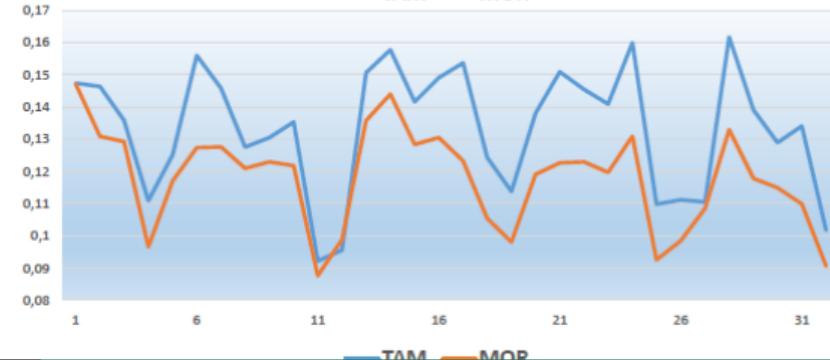
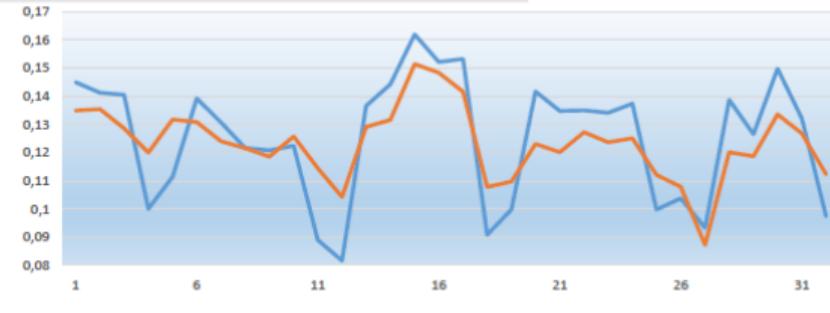
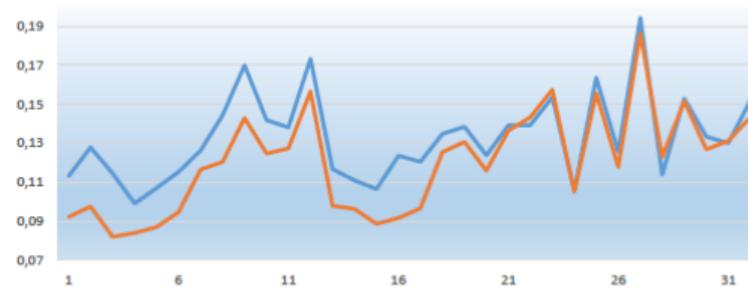
gdzie oznaczenia są takie same jak opisane w pracy Betlehem (2010).

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 (||)



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/ogladalnosc-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-ogladalnosci-z-danymi-od-netii-jacek-kurski-to->
- Domy mediowe: badanie oglądalności od TVP niemiarodajne, nie zastąpi pomiaru Nielsena <https://www.wirtualnemedia.pl/artykul/nowe-badanie-ogladalnosci-od-tvp-i-netii-domy-mediowe-jest-niemiarodajne-nie->

Przykład – CBOP

Tabela 3: Odsetek wakatów oferowanych na jedną zmianę w CBOP

Wielkość	\bar{y}_h	n_h	N_h	W_h
Małe (do 9)	0.7596	6 769	128 940	0.4381
Średnie (10-49)	0.7021	6 892	71 230	0.2420
Duże (50 i więcej)	0.2978	14 783	94 124	0.3198

gdzie: \bar{y}_h to odsetek wakatów oferowanych na jedną zmianę, n_h wielkość próby, N_h wielkości populacji, a $W_h = N_h/N$ to waga post-stratyfikacyjna. Naiwny estymator na podstawie tych danych wynosi: 0.5060 (około 51%).

Na podstawie tych danych możemy wyznaczyć estymator \bar{y}_{PS} :

$$\bar{y}_{PS} = 0.4381 \times 0.7596 + 0.2420 \times 0.7021 + 0.3198 \times 0.2978 = 0.5980 \quad (23)$$

Co by oznaczało, że około 60% wakatów w Polsce to wakaty oferowane na jedną zmianę.

Post-stratyfikacja – własności

Wariancja estymatora post-stratyfikacyjnego nie jest prosta do obliczenia ponieważ mamy dwa źródła zmienności:

- nieznany mechanizm tworzenia próby nielosowej (auto-selekcja, nieznane pokrycie populacji),
- znany mechanizm szacunku wartości globalnych (rejestr vs próba losowa).

Estymatory kalibracyjne

Calibration Estimators in Survey Sampling

JEAN-CLAUDE DEVILLE and CARL-ERIK SÄRNDAL*

This article investigates estimation of finite population totals in the presence of univariate or multivariate auxiliary information. Estimation is equivalent to attaching weights to the survey data. We focus attention on the several weighting systems that can be associated with a given amount of auxiliary information and derive a weighting system with the aid of a distance measure and a set of calibration equations. We briefly mention an application to the case in which the information consists of known marginal counts in a two- or multi-way table, known as *generalized raking*. The general regression estimator (GREG) was conceived with multivariate auxiliary information in mind. Ordinarily, this estimator is justified by a regression relationship between the study variable y and the auxiliary vector x . But we note that the GREG can be derived by a different route by focusing instead on the weights. The ordinary sampling weights of the k th observation is $1/\pi_k$, where π_k is the inclusion probability of k . We show that the weights implied by the GREG are as close as possible, according to a given distance measure, to the $1/\pi_k$ while respecting side conditions called *calibration equations*. These state that the sample sum of the weighted auxiliary variable values must equal the known population total for that auxiliary variable. That is, the calibrated weights must give perfect estimates when applied to each auxiliary variable. That is a consistency check that appeals to many practitioners, because a strong correlation between the auxiliary variables and the study variable means that the weights that perform well for the auxiliary variable also should perform well for the study variable. The GREG uses the auxiliary information efficiently, so the estimates are precise; however, the individual weights are not always without reproach. For example, negative weights can occur, and in some applications this does not make sense. It is natural to seek the root of the dissatisfaction in the underlying distance measure. Consequently, we allow alternative distance measures that satisfy only a set of minimal requirements. Each distance measure leads, via the calibration equations, to a specific weighting system and thereby to a new estimator. These estimators form a family of *calibration estimators*. We show that the GREG is a first approximation to all other members of the family; all are asymptotically equivalent to the GREG, and the variance estimator already known for the GREG is recommended for use in any other member of the family. Numerical features of the weights and ease of computation become more than anything else the bases for choosing between the estimators. The reasoning is applied to calibration on known marginals of a two-way frequency table. Our family of distance measures leads in this case to a family of *generalized raking procedures*.

Estymatory kalibracyjne (w j. polskim)

Marcin Szymkowiak

Podejście kalibracyjne w badaniach społeczno-ekonomicznych

Kalibracja dla prób nielosowych – idea

	X	W	Y	W*
Próba losowa				
Próba nielosowa				Ostatecznie, do wnioskowania, korzystamy tylko z próby nielosowej

Chcemy utworzyć wektor w^* tak, aby odtwarzał znane (lub estymowane) wartości globalne cech X .

Kalibracja – idea

(C1) Minimalizujemy funkcję odległości:

$$D(\mathbf{w}, \mathbf{d}) = \sum_{i=1}^m d_i G\left(\frac{w_i}{d_i}\right) \rightarrow \min \quad (24)$$

(C2) aby spełnione były tzw. równania kalibracyjne

$$\sum_{i=1}^m w_i x_{ij} = \mathbf{X}_j, \quad j=1, \dots, k, \quad (25)$$

(C3) oraz ograniczenia na wagę: $L \leq \frac{w_i}{d_i} \leq U$, where $0 \leq L \leq 1 \leq U$, $i=1, \dots, m$.

W przypadku prób nielosowych zwykle przyjmujemy, że waga $d_i = N/n$, gdzie N to wielkość populacji, a n to liczba rekordów w próbie nielosowej.

Kalibracja

W praktyce najczęściej korzysta się z następujących funkcji odległości

$$G_1(x) = \frac{1}{2}(x-1)^2, \quad (26)$$

$$G_2(x) = \frac{(x-1)^2}{x}, \quad (27)$$

$$G_3(x) = x(\log x - 1) + 1, \quad (28)$$

$$G_4(x) = 2x - 4\sqrt{x} + 2, \quad (29)$$

$$G_5(x) = \frac{1}{2\alpha} \int_1^x \sinh \left[\alpha \left(t - \frac{1}{t} \right) \right] dt, \quad (30)$$

Kalibracja

W przypadku zastosowania funkcji (26) otrzymujemy wektor wag w postaci analitycznej (tj. mamy konkretny wzór na ich wyznaczenie):

$$w_i^* = d_i + d_i \left(\mathbf{X} - \sum_{i \in S_n} d_i \mathbf{x}_i \right)^T \left(\sum_{i \in S_n} d_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \mathbf{x}_i, \quad (31)$$

gdzie \mathbf{X} to wektor znanych (lub estymowanych) wartości globalnych cech pomocniczych (np. cech demograficznych), $\sum_{i \in S_n} d_i \mathbf{x}_i$ możemy oznaczyć jako $\hat{\mathbf{X}}$,

W przypadku innych funkcji należy zastosować metody optymalizacyjne (por. Szymkowiak (2019), rozdział 2.3). W naszym przypadku stosować będziemy pakiet `survey` oraz funkcję `calibrate`.

Kalibracja

Ostatecznie, estymator kalibracyjnym wartości globalnej (np. liczby ofert pracy) dany jest wzorem

$$Y_{\text{cal}} = \sum_{i \in S_n} y_i w_i^*, \quad (32)$$

gdzie S_n oznacza próbę nielosową, a w_i^* to wektor pseudo-wag.

Podobnie, jak w przypadku post-stratyfikacji obliczenie wariancji tego estymatora nie jest proste.

Kalibracja a post-stratyfikacja

- Post-stratyfikacja jest szczególnym przypadkiem kalibracji.
- W kalibracji możemy stosować zarówno zmienne jakościowe, jak i ilościowe.
- W kalibracji możemy odtwarzać zarówno wartości globalne (np. liczebności), jak i średnie (np. średni dochód).
- Mamy większą kontrolę nad rozkładem wag w^* (przykładowo chcemy uniknąć wag ekstremalnych).

Kalibracja – krótki przykład #1

L.p.	Płeć	Zamieszkanie	Zatrudnienie	Wykształcenie
1	M	M	.	P
2	K	W	P	S
3	K	W	B	.
4	M	W	P	W
5	M	M	.	W
6	M	M	B	W
7	K	M	B	W
8	K	W	P	W
9	K	M	P	S
10	M	M	P	S
11	K	W	B	S
12	M	W	P	W
13	K	M	P	P
14	M	W	P	.
15	M	M	B	S
16	K	W	P	S
17	K	M	B	W
18	M	W	B	P
19	M	M	B	.
20	K	M	P	S

Kalibracja – krótki przykład #2

- Założymy, że celem badania jest stworzenie tabeli ukazującej status zatrudnienia w zależności od wykształcenia osoby. Ze względu na występujące w rejestrze braki danych uzyskana tabela nie będzie odpowiednia.
- Opis zmiennych (płeć: M-mężczyzna, K-kobieta; Zamieszkanie: M-miasto, W-wieś; Zatrudnienie: P-osoba pracująca, B-osoba bezrobotna; Wykształcenie: P-podstawowe, S-średnie, W-wyższe)

Wykształcenie	Status zatrudnienia		
	Pracująca	Bezrobotna	Razem
Wyższe	3	3	6
Średnie	5	2	7
Podstawowe	1	1	2
Razem	9	6	15

Algorytm wyznaczania wag kalibracyjnych – krok 1

- Wyjściowe wagi przypisujemy sztucznie w ten sposób, że dla jednostki dla której nie jest znana wartość co najmniej jednej z interesujących nas cech $d_i = 0$. Z kolei gdy znane są wartości wszystkich cech, w oparciu o które tworzona będzie tabela przyjmujemy $d_i = 1$.

L.p.	Płeć	Zamieszkanie	Zatrudnienie	Wykształcenie	d_i
1	M	M	.	P	0
2	K	W	P	S	1
3	K	W	B	.	0
4	M	W	P	W	1
5	M	M	.	W	0
6	M	M	B	W	1
7	K	M	B	W	1
8	K	W	P	W	1
9	K	M	P	S	1
10	M	M	P	S	1
...

Algorytm wyznaczania wag kalibracyjnych – krok 2

- W kolejnym kroku dokonujemy wyboru zmiennych, dla których znane są wartości dla wszystkich jednostek w rejestrze. Ponieważ informacja o płci osoby jak i jej miejscu zamieszkania znana jest dla wszystkich osób, zmienne te można wykorzystać celem ustalenia nowych wag w_i . Na potrzeby przykładu przyjęte zostały trzy zmienne pomocnicze: x_{i1} , x_{i2} , x_{i3} .

$$x_{i1} = \begin{cases} 1 & \text{jeżeli } i\text{-ta osoba jest kobietą,} \\ 0 & \text{jeżeli } i\text{-ta osoba jest mężczyzną,} \end{cases} \quad (33)$$

$$x_{i2} = \begin{cases} 1 & \text{jeżeli } i\text{-ta osoba jest mężczyzną,} \\ 0 & \text{jeżeli } i\text{-ta osoba jest kobietą,} \end{cases} \quad (34)$$

$$x_{i3} = \begin{cases} 1 & \text{jeżeli } i\text{-ta osoba mieszka w mieście,} \\ 0 & \text{jeżeli } i\text{-ta osoba mieszka na wsi.} \end{cases} \quad (35)$$

L.p.	Płeć	Zamieskanie	Zatrudnienie	Wykształcenie	d_i	x_{i1}	x_{i2}	x_{i3}
1	M	M	.	P	0	0	1	1
2	K	W	P	S	1	1	0	0
3	K	W	B	.	0	1	0	0
4	M	W	P	W	1	0	1	0
...

Algorytm wyznaczania wag kalibracyjnych – krok 3

- Tworzymy wektor złożony z wartości globalnych wszystkich zmiennych pomocniczych \mathbf{X} oraz wektor oszacowanych wartości globalnych $\hat{\mathbf{X}}$.
- W naszym przykładzie mamy: $\mathbf{X} = (10, 10, 11)^T$, $\hat{\mathbf{X}} = (9, 6, 8)^T$.
- Następnie wyznaczamy wagi kalibracyjne w_i korzystając ze wzoru (31).

Metody quasi-randomizacyjne

Wagi kalibracyjne w_i

L.p.	Płeć	Zamieszkanie	Zatrudnienie	Wykształcenie	d_i	\mathbf{x}_{i1}	\mathbf{x}_{i2}	\mathbf{x}_{i3}	w_i
1	M	M	.	P	0	0	1	1	0
2	K	W	P	S	1	1	0	0	1,0447761
3	K	W	B	.	0	1	0	0	0
4	M	W	P	W	1	0	1	0	1,6069652
5	M	M	.	W	0	0	1	1	0
6	M	M	B	W	1	0	1	1	1,7263682
7	K	M	B	W	1	1	0	1	1,1641791
8	K	W	P	W	1	1	0	0	1,0447761
9	K	M	P	S	1	1	0	1	1,1641791
10	M	M	P	S	1	0	1	1	1,7263682
11	K	W	B	S	1	1	0	0	1,0447761
12	M	W	P	W	1	0	1	0	1,6069652
13	K	M	P	P	1	1	0	1	1,1641791
14	M	W	P	.	0	0	1	0	0
15	M	M	B	S	1	0	1	1	1,7263682
16	K	W	P	S	1	1	0	0	1,0447761
17	K	M	B	W	1	1	0	1	1,1641791
18	M	W	B	P	1	0	1	0	1,6069652
19	M	M	B	.	0	0	1	1	0
20	K	M	P	S	1	1	0	1	1,1641791

Tworzenie tabelic – bez uwzględnienia wag kalibracyjnych

Wykształcenie	Status zatrudnienia		
	Pracująca	Bezrobotna	Razem
Wyższe	3	3	6
Średnie	5	2	7
Podstawowe	1	1	2
Razem	9	6	15

Tworzenie tabelic – z uwzględnieniem wag kalibracyjnych

Wykształcenie	Status zatrudnienia		
	Pracująca	Bezrobotna	Razem
Wyższe	4,27	4,05	8,32
Średnie	6,14	2,77	8,91
Podstawowe	1,16	1,61	2,77
Razem	11,57	8,43	20

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

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 Lee¹

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.

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

283

Vol. 48, No. 2, pp. 283-311

Statistics Canada, Catalogue No. 12-001-X

Statistical inference with non-probability survey samples

Changbao Wu¹

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;

Propensity score weighting – idea

	X1 (płeć)	X1 (wiek)	Y (słuchanie podcastów)	w (waga)	R	P(R=1 X1, X2)
Próba losowa	M	34	x	4	0	0,65
	M	32	x	2	0	0,85
	K	50	x	5	0	0,35
	K	40	x	10	0	0,23
Próba niełosowa	M	32	Tak	?	1	0,85
	M	34	Nie	?	1	0,65
	K	40	Tak	?	1	0,23
	K	50	Tak	?	1	0,35

w* = 1 / P(waga finalna)
x
x
x
x
1,1765
1,5385
4,3478
2,8571

Ten zbiór wykorzystamy do estymacji

- Mamy dwa zbiory danych (próba losowa i niełosowa).
- Mamy dwie wspólne zmienne (X_1, X_2) oraz jedną obserwowaną tylko w próbie niełosowej (Y).
- W próbie losowej mamy wagi (d).
- Szacujemy prawdopodobieństwo $P(R = 1)$ uwzględniając wspólne zmienne (X_1, X_2).
- Dla jednostek z próby niełosowej przypisujemy wagi $w^* = 1/P(R = 1|X_1, X_2)$.

Propensity score – założenia (cz. 1)

- Różne nazwy na to samo: propensity score adjustment (PSA), propensity score weighting (PSW), inverse probability weighting (IPW).
- W metodzie *propensity score* zakładamy, że
 - ① potrafimy rozróżnić jednostki, które są i nie są w próbie nielosowej,
 - ② dysponujemy źródłem, które zawiera jednostki nie występujące w źródle niełosowym.
- Przykład:
 - Źródło big data oraz dane dotyczące całej populacji
 - Źródło big data oraz próba losowa, w której znajdują się również jednostki obserwowane w big data.

Propensity score – oznaczenia

Podstawowe oznaczenia

- Niech $R_i = \{0, 1\}$ oznacza zmienną określającą przynależność do próby niełosowej ($R_i = 1$).
- Niech zmienne \mathbf{x}_i oznacza wektor zmiennych pomocniczych, obserwowanych w obydwu źródłach danych.
- Niech $\pi_i = P(R_i = 1 | \mathbf{x}_i)$ prawdopodobieństwo przynależności do danego źródła.
- π_i nie jest znane jest estymowane na podstawie danych.

Propensity score – założenia (cz. 2)

Formalnie, założenia metody *propensity score* są następujące:

- Zmienna selekcji / inkluzji R_i oraz badana przez nas cecha y_i są warunkowo niezależne gdy pod uwagę weźmiemy cechy \mathbf{x}_i . Innymi słowy $\pi_i = P(R_i = 1|y, \mathbf{x}_i) = P(R_i = 1|\mathbf{x}_i)$ – **Inaczej:** dobór jest nieinformatywny (w literaturze przedmiotu: missing at random, ignorable).
- Wszystkie jednostki w populacji mają niezerowe prawdopodobieństwo inkluzji do próby nieosowej ($\pi_i > 0$) – **Inaczej:** brak błędów pokrycia.
- Zmienne R_i oraz R_j są niezależne gdy uwzględnimy \mathbf{x}_i dla $i \neq j$ – **Inaczej:** obserwacje są niezależne (m.in. brak duplikatów lub jakichś zmiennych \mathbf{z}_i , których nie obserwujemy).

Propensity score weighting

π_i możemy estymować wiele różnych sposobów. Na przykład, możemy do tego celu wykorzystać

- model regresji logistycznej (w tym np. LASSO)

$$\log \left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \right) = \mathbf{x}_i^T \boldsymbol{\beta} \quad (36)$$

- algorytmy uczenia maszynowego

$$P(R_i = 1 | \mathbf{x}_i) = f(\mathbf{x}_i), \quad (37)$$

gdzie $f(\cdot)$ jest dowolną funkcją.

Propensity score weighting – estymator

Po oszacowaniu π , wykorzystujemy następujący estymator, który oznaczamy jako IPW/PSW/PSA

$$\hat{\theta}_{IPW/PSW/PSA} = \frac{\sum_{i \in S_n} y_i \hat{\pi}_i^{-1}}{\sum_{i \in S_n} \hat{\pi}_i^{-1}}, \quad (38)$$

gdzie S_n oznacza próbę nielosową, π_i^{-1} to odwrotność sklonności. Uwaga: nie robimy tutaj sumy ponieważ suma π_i^{-1} nie daje nam populacji.

Estymacja wariancji tego estymatora nie jest taka prosta i zostanie pokrótce omówiona na kolejnym slajdzie.

Propensity score weighting – teoria

Ostatnie, najbardziej aktualne artykuły poświęcone wykorzystaniu metody propensity score weighting dla prób nielosowych:

- Kim, J. K., & Wang, Z. (2019). **Sampling techniques for big data analysis.** *International Statistical Review*, 87, S177-S191
 - estymacja tylko na podstawie próby losowej,
 - propozycja estymatora IPW/PSW/PSA oraz jego wariancji (dla prostych schematów losowania próby losowej).
- Chen, Y., Li, P., & Wu, C. (2020). **Doubly robust inference with nonprobability survey samples.** *Journal of the American Statistical Association*, 115(532), 2011-2021.
 - estymacja IPW/PSW/PSA dla wszystkich obserwacji (oba źródła jednocześnie).
 - propozycja estymatora IPW/PSW/PSA oraz jego wariancji (ogólna postać dla wszelkiego rodzaju schematów losowania próby losowej).

Propensity score weighting – wady i zalety

Wśród zalet możemy wymienić:

- prosta idea i implementacja,
- jeden zestaw wag dla całego zbioru danych,
- w bardziej (powiedzmy) zaawansowanej wersji wymaga wyłącznie znajomości wartości globalnych / średnich dla cech x_i .

Wśród wad możemy wymienić:

- metoda działa **tylko wtedy gdy poprawnie określmy model** dla π (postać, zmienne),
- estymator PS nie jest efektywny,
- w podstawowej wersji wymaga danych jednostkowych (próba losowa i nielosowa),
- musimy zidentyfikować jednostki między źródłami,
- wagi ($w_i^* = 1/\hat{\pi}_i$) nie odtwarzają wartości globalnych z populacji (np. liczby kobiet, mężczyzn; charakterystyk podmiotów gospodarczych),
- jeden zestaw wag dla całego zbioru danych.

Podejście oparte na modelu

- W podejściu opartym na modelu zakładamy, że interesuje nas $E(Y|X)$,
- Zakładamy, że model $E(Y|X, R = 1) = E(Y|X) = \mu(y_i|\mathbf{x}_i)$,
- Budujemy model na próbie nielosowej (np. regresja liniowa, logistyczna) i aplikujemy na całą populację (ewentualnie próbę)
- Estymator, w przypadku gdy znamy całą populację na postać:

$$\hat{\theta}_M = \sum_{i \in S_A} y_i + \sum_{i \in U \setminus S_A} \hat{y}_i, \quad (39)$$

gdzie S_A to próba nielosowa, a $U \setminus S_A$ to pozostała część populacji.

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. Elliott¹

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

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

Podejście oparte na modelu

Combining Non-probability and Probability Survey Samples Through Mass Imputation

Jae Kwang Kim¹, Seho Park², Yilin Chen³, and Changbao Wu³

¹ Department of Statistics, Iowa State University,

² Department of Biostatistics, Indiana University School of Medicine,

³ 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.

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 nieosowej)	
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 nieosowa	M	32	Tak	?	1		
	M	34	Nie	?	1		
	K	40	Tak	?	1		
	K	50	Tak	?	1		

Rysunek 9: Podejście I: Przepisanie wartości z próby nieosowej

	X1 (płeć)	X1 (wiek)	Y (słucha 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 nieosowa	M	32	Tak	?	1		
	M	34	Nie	?	1		
	K	40	Tak	?	1		
	K	50	Tak	?	1		

Rysunek 10: Podejście II: wartości przewidywane z modelu zbudowanego na próbie nieosowej

Masowa imputacja – podejście I

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

- ① Dla próby nielosowej \mathcal{S}_A oraz losowej \mathcal{S}_B określamy zestaw wspólnych cech \mathbf{X} .
- ② Następnie, dla każdej jednostki $i \in \mathcal{S}_B$ szukamy najbliższej jednostki ze zbioru $k \in \mathcal{S}_A$, tak żeby

$$d(\mathbf{x}_k, \mathbf{x}_i) = \|\mathbf{x}_k - \mathbf{x}_i\| \quad (40)$$

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

- ③ Po znalezieniu odpowiednich sąsiadów, wyznaczamy estymator. Przykładowo, estymator wartości średniej $\theta = N^{-1} \sum_{i \in U} y_i$ dany będzie:

$$\hat{\theta}_{M1} = \sum_{i \in \mathcal{S}_B} w_i y_i / \sum_{i \in \mathcal{S}_B} w_i. \quad (41)$$

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:

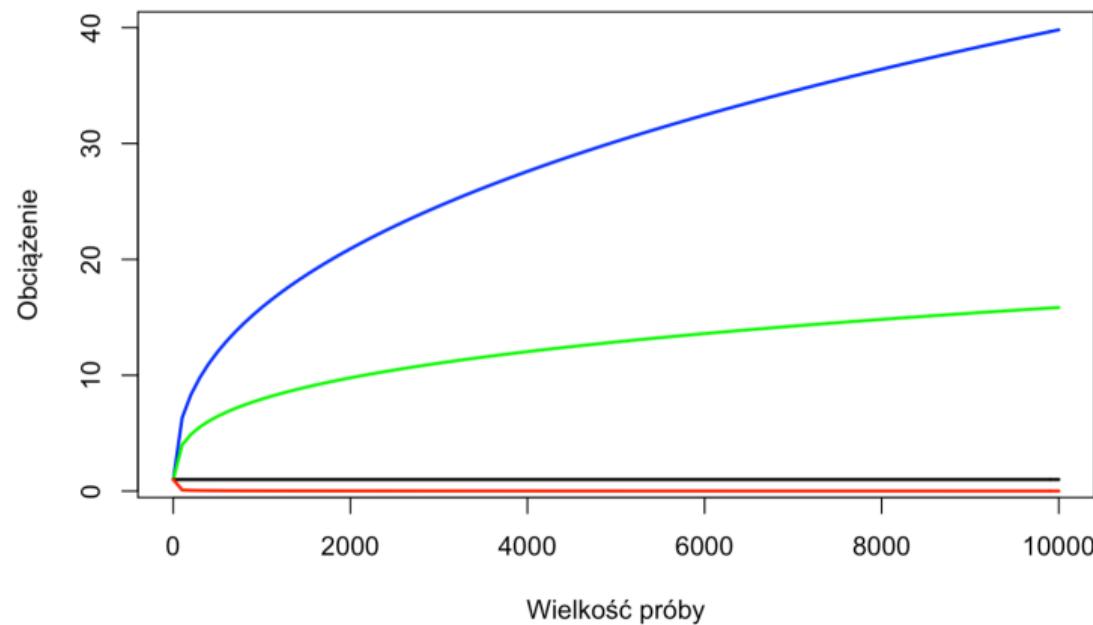
$$Bias(\hat{\theta}_{M1}) = O_p\left(n^{1/2 - 1/p}\right) \quad (42)$$

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 11: 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 nieosowej dla jednostek k z próby losowej należy wykorzystać na podstawie wyłącznie jednej zmiennej, stosuje się następujące podejście:

- ① Budujemy model $m(\mathbf{x}_i; \boldsymbol{\beta})$ na próbie nieosowej \mathcal{S}_A otrzymując $\hat{y}_i = m(\mathbf{x}_i; \hat{\boldsymbol{\beta}})$.
- ② Stosujemy model $m(\mathbf{x}; \hat{\boldsymbol{\beta}})$ na próbie losowej \mathcal{S}_B otrzymując $\hat{y}_k = m(\mathbf{x}_k; \hat{\boldsymbol{\beta}})$.
- ③ 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 .
- ④ Następnie wyznaczamy estymator

$$\hat{\theta}_{M1} = \sum_{i \in \mathcal{S}_B} w_i y_i / \sum_{i \in \mathcal{S}_B} w_i. \quad (43)$$

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

$$\hat{\theta}_{M2} = \sum_{i \in S_B} w_i \hat{y}_i / \sum_{i \in S_B} w_i. \quad (44)$$

- 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.

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

$$\hat{\theta}_{\text{DR}} = \underbrace{\frac{1}{N} \sum_{i=1}^N \frac{R_i \left\{ y_i - m(\hat{\beta}; \mathbf{x}_i) \right\}}{\rho(\hat{\lambda}; \mathbf{x}_i)}}_{\text{Średnia ważona reszt z modelu}} + \underbrace{\frac{1}{N} \sum_{i=1}^N m(\hat{\beta}; \mathbf{x}_i)}_{\text{Średnia z predykcji dla całej populacji}}, \quad (45)$$

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 – w telegraficznym skrócie

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

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

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

$$\hat{\theta}_{\text{DR1}} = \underbrace{\frac{1}{N} \sum_{i \in \mathcal{S}_A} w_i^* \left\{ y_i - m(\mathbf{x}_i, \hat{\beta}) \right\}}_{\text{Próba nielosowa}} + \underbrace{\frac{1}{N} \sum_{i \in \mathcal{S}_B} d_i m(\mathbf{x}_i, \hat{\beta})}_{\text{Próba losowa}}, \quad (46)$$

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

Podwójnie odporne estymatory

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

$$\hat{\theta}_{DR2} = \overbrace{\frac{1}{\hat{N}^A} \sum_{i \in S_A} w_i^* \left\{ y_i - m(\mathbf{x}_i, \hat{\beta}) \right\}}^{\text{Próba niełosowa}} + \underbrace{\frac{1}{\hat{N}^B} \sum_{i \in S_B} d_i m(\mathbf{x}_i, \hat{\beta})}_{\text{Próba losowa}}, \quad (47)$$

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 nieosowej ($n_B/N_A = o(1)$) 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 nieosowej 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_BNastępnie 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.

Tyle na dzisiaj, dziękuję!