

OPIS DOBORU PRÓBY

Opis doboru próby do badania sondażowego n. t. „Jak się w Polsce żyje młodym?” zrealizowanego w ramach projektu *„Między przeszłością a przyszłością. Dążenia i perspektywy życiowe pokolenia III RP – uwarunkowania środowiskowe, kulturalne i edukacyjne”*.

Spis treści

Założenia	2
Populacja	2
Operat	3
Podział administracyjny kraju	3
PESEL	3
Tablica DEMOGRAF	4
Schemat doboru	4
Warstwowanie	4
Alokacja próby między warstwy	4
Losowanie zespołów - elementarnych jednostek podziału terytorialnego	5
Losowanie osób	5
Wyniki losowania	5
Własności statystyczne próby wylosowanej	6
Wagi analityczne	6
Ważenie po-realizacyjne	6

Założenia

Populacja

Do wylosowania jest próba o liczności $n=3920$ wiązowana po 10 elementów na wiązkę. Ma być pobrana z populacji mieszkańców Polski w wieku 19-30 lat (ukończone).

Próba ma być podstawą minimalnie obciążonych oraz maksymalnie dokładnych oszacowań parametrów wspomnianej populacji.

Wedle PESEL w 2012 roku populacja ta liczyła około 6,5 mln osób a jej rozkład terytorialny oraz rozkład płci i wieku były następujące:

Tabela 1. Rozkład terytorialny populacji 19-30 lat

województwo	1 Wieś	2 Miasto do 50 tys.	3 Miasto 50 - 200 tys.	4 Miasto powyżej 200 tys.	razem
2 dolnośląskie	153700	154633	83215	88914	480462
4 kujawsko-pomorskie	153098	72221	80508	55675	361502
6 lubelskie	210594	84794	44979	54256	394623
8 lubuskie	69381	73108	38258	0	180747
10 łódzkie	156873	97557	55674	100538	410642
12 małopolskie	259357	168192	35427	105718	568694
14 mazowieckie	327526	183794	60487	256270	828077
16 opolskie	86527	60781	28740	0	176048
18 podkarpackie	228829	105266	65755	0	399850
20 podlaskie	86024	57383	26180	48698	218285
22 pomorskie	139327	118072	26238	101008	384645
24 śląskie	166146	134793	91851	375164	767954
26 świętokrzyskie	126758	43893	54965	0	225616
28 warmińsko-mazurskie	116583	93922	59693	0	270198
30 wielkopolskie	275073	166745	78301	81700	601819
32 zachodniopomorskie	100574	102875	28145	57740	289334
razem	2656370	1718029	858416	1325681	6558496

Tabela 2. Rozkład płci i wieku w populacji 19-30 lat.

Nr kategorii	Kategoria wiek	Mężczyzna	Kobieta	razem
1	19-20	547134	522821	1069955
2	21-22	561955	542638	1104593
3	23-24	590434	574201	1164635
4	25-26	648212	633177	1281389
5	27-28	665680	651603	1317283
6	29-30	313398	307243	620641
	razem	3326813	3231683	6558496

Operat

Podział administracyjny kraju

Najniższą jednostką podziału administracyjnego kraju (nazywaną NUTS 5 albo NTS 5) jest gmina. Jest ich w Polsce 2498. prowadzonym przez GUS W systemie TERC każda gmina ma unikalny identyfikator zbudowany wedle następującej zasady:

1-2	3-4	5-6	7
XX	XX	XX	X
Województwo	Powiat	Gmina	Typ gminy

Ostatnia cyfra identyfikatora służy także do wyodrębnienia z obszaru gmin mieszanych ich części miejskiej i wiejskiej a z obszaru dużych miast ich dzielnic:

7 cyfra	znaczenie
1	gmina miejska,
2	gmina wiejska,
3	gmina miejsko-wiejska,
4	miasto w gminie miejsko-wiejskiej,
5	obszar wiejski w gminie miejsko-wiejskiej,
8	dzielnice gminy Warszawa-Centrum,
9	delegatury i dzielnice innych gmin miejskich

Dla przykładu gminy powiatu białogardzkiego (kod powiatu 3201) leżącego w województwie zachodnio pomorskim (kod 32) mają następujące kody TERC:

- 3201011 - miasto Białogard - gmina miejska,
- 3201022 - gmina wiejska Białogard,
- 3201033 - gmina miejsko-wiejska Karlino,
- 3201034 - miasto Karlino w gminie miejsko-wiejskiej,
- 3201035 - obszar wiejski w gminie miejsko-wiejskiej Karlino

Po uwzględnieniu podziału gmin mieszanych (wiejsk-miejskich) liczba **elementarnych** jednostek podziału administracyjnego - obszarów posiadających unikalny 7-cyfrowy identyfikator (nazywany dalej kodem terytorialnym, KT) - wynosi 3082¹. Każda miejscowość stanowiąca część adresu zamieszkania (zameldowania) ma jeden z 3082 kodów terytorialnych. Kody te są stosowane w systemie ewidencji ludności PESEL.

PESEL

Baza PESEL zawiera dane osobowe oraz adresy zameldowania wszystkich obywateli Polski. Ma stosunkowo niewielki odsetek (2%-3%) zapisów błędnych (błędne adresy lub dane osób zmarłych).

PESEL jest w praktyce sondażowej operatem niewygodnym, gdyż odsetek osób niedostępnych w miejscu ich zameldowania dla niektórych regionów przekracza 20%. Oznacza to że przekroczenie response rate powyżej 75% jest dla próby pobranej z PESEL praktycznie niemożliwe.

¹ Zarówno liczba gmin jak i elementarnych obszarów ulega niewielkim zmianom. Dane cytowane pochodzą z roku 2012.

Operat ten ma jednak ważną zaletę – łatwość i wyoka rzetelność terenowej kontroli ankieterów realizujących badanie, pozwala także na użycie poziomu response rate jako parametru jakości realizacji uwzględnianego w kontrakcie z wykonawcą badania. Te zalety przesądziły o jego wyborze w losowaniu próby do badania².

W pierwszym stopniu losowania wykorzystano wyznaczoną w PESEL na potrzeby zespołu badawczego statystykę ludności Polski, w drugim pracownicy PESEL korzystali bezpośrednio z bazy danych systemu.

Tablica DEMOGRAF

Statystyka ludności Polski została utworzona w PESEL na podstawie stanu jego bazy z września 2012 roku jest tablicą o nazwie DEMOGRAF. Tablica ta zawiera tyle wierszy ile jest **różnych** 7-cyfrowych kodów terytorialnych oraz tyle kolumn, ile kategorii płci i wieku jest niezbędnych do alokacji próby. W komórce tablicy dla każdej kategorii wieku i płci znajduje się liczba osób zgłoszonych na obszarze odpowiadającym elementarnej jednostce podziału terytorialnego. Z formalnego punktu widzenia DEMOGRAF jest łącznym rozkładem płci, wieku i miejsca zamieszkania rozumianego jako ten obszar.

Kod terytorialny elementarnej jednostki podziału terytorialnego zawiera informację o województwie, powiecie oraz typie obszaru (wiejski lub miejski). Zsumowanie wszystkich komórek z tego samego wiersza tablicy daje w rezultacie liczebność populacji zameldowanych na terenie odpowiadającym kodowi terytorialnemu.

Agregacja wybranych wierszy i kolumn tablicy pozwala na wyznaczenie populacyjnych rozkładów płci, wieku i miejsca zamieszkania (patrz tablice 1 i 2 powyżej) a w konsekwencji wyznaczenie liczebności warstw stosowanych w doborze próby. W konsekwencji prawdopodobieństwa inkluzji można wyznaczyć dla każdego elementu populacji także w złożonych schematach losowania.

Schemat doboru

Zastosowano losowanie dwustopniowe:

1. W pierwszym etapie, niezależnie z każdej z warstw losowano ze zwracaniem jednostki losowania pierwszego stopnia – zespoły odpowiadające elementarnym jednostkom podziału terytorialnego kraju. Alokacja próby między warstwy była nieproporcjonalna i nierównomierna. Warstwowanie objęło całe terytorium Polski.
2. W drugim etapie z każdego z wylosowanych zespołów losowano bezzwrotnie metodą doboru systematycznego jednostki losowania drugiego stopnia - respondentów, przy czym liczba tych jednostek była taka sama dla każdego zespołu i wynosiła 10.

Warstwowanie

Warstwy zdefiniowano przy użyciu podziału terytorialnego kraju na 16 województw a w każdym z nich obszary odpowiadające elementarnym jednostkom podziału terytorialnego kraju podzielono na 4 klasy wielkości określone przez wielkość populacji generalnej zamieszkującej obszar. W populacji generalnej było 59 niepustych warstw (patrz tabela 1). Liczba osób w wieku 19-30 lat w warstwach wahała się między 26 a 375 tysięcy.

Alokacja próby między warstwy

Zastosowano alokację nieproporcjonalną z nadreprezentacją mieszkańców Warszawy. Rozkład 392 zespołów między warstwy terytorialne przedstawia tabela 3.

Tabela 3. Alokacja zespołów między warstwy terytorialne.

województwo	1 Wieś	2 Miasto do 50 tys.	3 Miasto 50 - 200 tys.	4 Miasto powyżej 200 tys.	razem
2 dolnośląskie	9	9	6	5	29
4 kujawsko-pomorskie	10	4	6	2	22
6 lubelskie	13	6	2	2	23
8 lubuskie	3	5	3	0	11
10 łódzkie	6	5	3	5	19

² W dodatku dla publicznych szkół wyższych koszt losowania próby z operatu PESEL jest niewielki.

12 małopolskie	10	9	2	6	27
14 mazowieckie	10	6	3	60	79
16 opolskie	5	3	3	0	11
18 podkarpackie	10	6	4	0	20
20 podlaskie	6	3	2	2	13
22 pomorskie	10	9	2	6	27
24 śląskie	7	6	4	24	41
26 świętokrzyskie	6	3	3	0	12
28 warmińsko-mazurskie	6	5	3	0	14
30 wielkopolskie	12	9	5	5	31
32 zachodniopomorskie	4	4	2	3	13
razem	127	92	53	120	392

Nieproporcjonalność alokacji dotyczy praktycznie wyłącznie województwa mazowieckiego, w którym lekko niedoreprezentowana jest populacja mieszkańców wsi (-2,4% liczby zespołów) zaś nadreprezentowani są mieszkańcy Warszawy (+11,4%). Przy estymacji parametrów całej populacji niezbędne będzie zastosowanie wag analitycznych wynikających z tej nieproporcjonalności oraz jej zakłóceń wynikających z niezbędnych okrągłości (wynoszą one od -1,4% do +0,5% liczby zespołów).

Losowanie zespołów - elementarnych jednostek podziału terytorialnego

Zespoły losowano ze zwracaniem z prawdopodobieństwami proporcjonalnymi do liczby ich mieszkańców w wieku 19-30 lat. Oznacza to, że niektóre zespoły zostały wylosowane wielokrotnie – na przykład dzielnica Warszawa-Mokotów została wylosowana 12-krotnie, co oznacza że z tej dzielnicy zostanie wylosowanych 120 jej mieszkańców w wieku 19-30 lat. Liczba różnych zespołów wylosowanych wynosi 161.

Losowanie osób

Liczba powtórzeń losowania zespołu definiuje liczbę osób, które będą z niego losowane. Przy znanej z tablicy DEMOGRAF liczbie ludności w wieku 19-30 zamieszkałej na terenie odpowiadającym wylosowanej jednostce podziału administracyjnego znane jest zatem warunkowe prawdopodobieństwo inluzji dla jednostek populacji zamieszkałej na tym obszarze.

Listę 161 zespołów wraz z liczbą osób, które należy z nich wylosować przekazano do PESEL, z prośbą o przeprowadzenie drugiego stopnia losowania. Uzgodniono, że losowanie dla każdego zespołu zostanie przeprowadzone metodą doboru systematycznego z losowym wyborem punktów startowych.

Wyniki losowania

Próba otrzymana z PESEL zawierała imię, nazwisko, datę urodzenia, płeć i adres zameldowania wylosowanej osoby wraz z kodem terytorialnym miejscowości zamieszkania. Pozwoliło to sprawdzić, czy losowanie przeprowadzono zgodnie z zamówieniem i wyznaczyć wagi analityczne niezbędne w analizie danych z próby przed jej wydaniem do realizacji wykonawcy terenowej części badania..

Z tablicy DEMOGRAF wyznaczono łączny populacyjny rozkład czterech zmiennych:

1. Województwa zamieszkania (16 województw)
2. Klasy wielkości miejscowości zamieszkania (4 klasy),
3. Płci (2 kategorie)
4. Wiek (6 przedziałów)

Rozkład ten ma 708 komórek ($59 * 2 * 6 = 708$) a odpowiadających w zbiorze danych wartościom zmiennej RKPA utworzonej w następujący sposób:

$$WKPA = 1000 * woj + 100 * klwm4 + 10 * Plec + AgCat$$

Gdzie

- „woj” - nr województwa [2-32],
- „klwm4” klasa wielkości miejscowości [1-4]
- „Plec” płeć [1-2]
- „AgCat” przedział wieku [1-6]: 1. 19-20; 2. 21-22; 3. 23-24; 4. 25-26; 5. 27-28; 6. 29-30

W próbie wylosowanej (n=3920) zmienna WKPA przyjmuje 632 wartości a w próbie zrealizowanej - 484. Populacyjne i próbkowe częstości tej zmiennej posłużą do wyznaczenia wag analitycznych metodą uproszczoną.

Własności statystyczne próby wylosowanej

Próba została wylosowana w sposób złożony a jej alokacja była nieproporcjonalna, wszystkie populacyjne oszacowania wymagają więc uwzględnienia wag analitycznych.

Ponadto należy zbadać wielkość wpływu (zazwyczaj pogorszenia) schematu doboru próby na dokładność populacyjnych oszacowań (DEFF - design effect).

Oba powyższe zadania, wyznaczenie wag oraz DEFF wymagają specjalistycznego oprogramowania (Uniwersytet Warszawski nie zakupił modułu SPSS Complex Samples) i zostały zaplanowane na następny etap badania. Teraz zajmiemy się tylko wyznaczeniem wag analitycznych przy użyciu uproszczonej metody wykorzystującej zmienną WKPA, której jedynym zadaniem jest neutralizacja wpływu nieproporcjonalności alokacji na obciążenie populacyjnych estymatorów.

Wagi analityczne

Waga analityczna $wagaAA$ korygująca nieproporcjonalność alokacji próby została jest proporcjonalna do ilorazu względnych częstości warstwy WKPA w próbie i w populacji:

$$wagaAA \propto \frac{n_{WKPA} / n}{N_{WKPA} / N}$$

gdzie: n_{WKPA} jest liczbą elementów warstwy WKPA w wylosowanej próbie o liczebności n
 N_{WKPA} jest liczbą elementów warstwy WKPA w populacji o liczebności N .

Po uwzględnieniu wagi $wagaA$ estymatory populacyjnych parametrów stają się nieobciążone. Nietety, ich wariancje rosną ze względu na stosunkowo dużą wariancję wielkości wag:

Tabela 4. Własności wagi analitycznej $wagaAA$:

	Minimum	Maximum	Średnia	Odch standardowe
$wagaAA$	,235	6,229	1,000	,788

Ważenie po-realizacyjne

Spośród 708 możliwych komórek rozkładu zmiennej WKPA w populacji w zrealizowanej próbie znalazło się 484. Dla każdej z nich wyznaczono wagę po-realizacyjną $wagaAAR$ korygującą nieproporcjonalną alokację zespołów. Waga ta jest proporcjonalna do ilorazu sumy $wagaAA$ w warstwie WKPA w próbie zrealizowanej oraz populacyjnej liczebności tej warstwy:

$$wagaAAR \propto \frac{n(wagaAA)_{WKPA} / n_R}{N_{WKPA} / \sum_{WKPA} N_{WKPA}}$$

gdzie:

$n(wagaAA)_{WKPA}$ jest sumą wag analitycznych ($wagaAA$) przypisanych elementom z próby **wylosowanej** $n=3920$

N_{WKPA} jest liczbą elementów warstwy WKPA w populacji zaś sumowanie tych liczebności obejmuje tylko te warstwy WKPA, które występują w próbie **zrealizowanej** o liczebności $n_R=1620$

Użycie powyższych wag zmniejsza obciążenie estymatorów populacyjnych spowodowane nieproporcjonalną alokacją próby zespołów, jest jednak zarazem dodatkowym czynnikiem zwiększającym ich wariancję, gdy stopień realizacji próby (response-rate) w warstwach WKPA nie jest identyczny.

Tabela 5. Własności wagi po-realizacyjnej $wagaAAR$:

	Minimum	Maximum	Średnia	Odch standardowe
$wagaAAR$	,039	21,720	1,000	2,002