Metoda reprezentacyjna

Łukasz Wawrowski

Contents

List of Tables

List of Figures

Literatura

- Heeringa, S. G., West, B. T., & Berglund, P. A. (2017). Applied survey data analysis. CRC press.
- Lumley, T. (2011). Complex surveys: a guide to analysis using R. John Wiley & Sons.
- Falissard, B. (2011). Analysis of questionnaire data with R. CRC Press.
- De Leeuw, E. D., Hox, J. J., & Dillman, D. A. (2008). International handbook of survey methodology. Taylor & Francis Group.

Chapter 1

Wprowadzenie do metody reprezentacyjnej

Prezentacja

1.1 Metoda reprezentacyjna

Badania statystyczne stanowią podstawę funkcjonowania państwa oraz społeczeństwa - dzięki nim znany jest poziom inflacji czy bezrobocia. W praktyce dominują badania oparte na próbie, ponieważ badanie wszystkich jednostek często jest utrudnione oraz nieopłacalne. Dzięki zastosowaniu odpowiednich metod statystycznych wyniki zebrane na podstawie próby można z powodzeniem uogólniać na całą populację. Oczywiście znając wielkość błędu jaki się wówczas popełnia.

Metoda reprezentacyjna ma na celu określenie zasad dotyczących projektowania, zbierania, przetwarzania oraz analizy danych, które wpływają na koszt i jakość badania. Innymi słowy metoda reprezentacyjna zajmuje się metodologią badań statystycznych.

W Polsce prym w prowadzeniu badań wiedzie Główny Urząd Statystyczny, który w Programie Badań Statystycznych Statystyki Publicznej publikuje listę prowadzonych badań. Oprócz tego na rynku działa wiele firm, które zajmują się badaniami statystycznymi. Wśród najpopularniejszych można wskazać CBOS, Kantar Millward Brown i wiele innych. Najgłośniej o tych podmiotach mówi się przy okazji wyborów, kiedy na tapetę brane są sondaże. Z dobrodziejstw badań statystycznych korzystają także prywatne przedsiębiorstwa produkcyjne, które z wykorzystaniem tych metod prowadzą badania jakościowe.

Projektując badanie statystyczne musimy zdefiniować następujące charakterystyki badania:

1. Cel badania
2. **Populacja generalna** którą badanie ma opisywać
3. Źródła z których może zostać wylosowana próba - **operat losowania**
4. Sposób w jaki próba zostanie wylosowana - **schemat losowania**
5. Sposób zbierania danych (CATI, CAWI, CAPI I PAPI)

Badanie, które przeprowadzimy zgodnie z powyższą listą cechuje się tym, że jest oparte na **próbie losowej**. Oznacza to, że w sporym uproszczeniu, wyniki badania przeprowadzonego na próbie losowej można uogólniać na populację generalną.

Wśród zalet tego podejścia można wskazać przede wszystkim znajomość wielkości błędu jaki popełniamy przy tym uogólnieniu. Ponadto można wskazać proste zależności pomiędzy liczebnością próby, błędem badania oraz kosztami. Większa liczebność próby implikuje mniejszy błąd badania, natomiast wiąże się ze zwiększeniem kosztów badania. Celem autora badania powinno być znalezienie kompromisu pomiędzy budżetem dostępnym na przeprowadzenie badania a błędem badania.

Do przeprowadzania takiego badania wymagana jest jednak wiedza dziedzinowa oraz dostęp do operatu losowania (czasami jest to poważny problem).

Wcielimy się na chwilę w rolę studentów koła naukowego, które chce przeprowadzić badanie studentów Uniwersytetu Ekonomicznego w Poznaniu na temat wrażeń z ostatniej sesji. Wiadomo, że było super, ale wartoby mieć jakieś liczby, które to potwierdzą.

1. Cel badania: poznanie opinii studentów na wybrany temat.
2. Populacja generalna: wszyscy studenci UEP (8415 osób w roku 2018/2019)

Zakładając, że w kole naukowym mamy 10 osób to dotarcie do wszystkich studentów i skłonienie ich do wypełnienia ankiety byłoby problematyczne. Zatem decydujemy się na przeprowadzenia badania na próbie.

3. Operat losowania: lista wszystkich studentów UEP wraz z danymi kontaktowymi

W przypadku studentów pozyskanie listy z Biura Obsługi Studentów wraz z dodatkowymi informacji nie powinno być problemem.

4. Schemat losowania: wylosowanie z operatu np. 5% próby z uwzględnieniem płci i kierunku studiów

Najważniejszy moment czyli wybranie osób, które wypełnią naszą ankietę. 5% z 8415 to około 420 studentów, więc każdy członek koła przeprowadzi wywiad z około 40 studentami - jest to do zrobienia. Powinien być to dobór losowy, ale uwzględniający strukturę np. płci i kierunku studiów. Zależy nam na tym, żeby próba wylosowana do badania była miniaturą populacji.

5. Zebranie danych: ankieta w Google Forms

Wysłanie ankiety poprzez e-mail może spowodować, że wiele osób w ogóle nie kliknie w załączony link. Dużo efektywniejszą formą zbierania danych będzie CAPI - wywiad osobisty z wykorzystaniem formularza internetowego.

1.2 Rys historyczny

Świat

1. Booth, C., (1889-1903), Life and Labour of the People of London

Zebranie danych o ubóstwie dla każdego domu w Londynie.

2. Thurstone, L. i Chave, E., (1929), The Measurement of Attitude, Chicago: University of Chicago. Likert, R., (1932), A technique for the measurement of attitudes, Archives of psychology, 140, pp. 5-53.

Wielki kryzys i ograniczone przez to środki przyspieszają rozwój metody reprezentacyjnej. Opracowanie przez Rensisa Likerta pięciostopniowej skali.

3. Hansen, M., (1939), Survey of Unemployment.

Pierwsze poważne badanie dotyczące bezrobocia.

4. Deming, W., (1950), Some Theory of Sampling, New York: Dover. Hansen, M., Hurwitz, W. i Madow, W., (1953), Sample survey methods and theory, Wiley.

Pierwsze podręczniki dotyczące metody reprezentacyjnej.

Polska

Neyman, J., (1933), Zarys teorii i praktyki badania struktury ludności metodą reprezentacyjną

Zasępa, R., (1972), Metoda reprezentacyjna

Bracha, Cz. (1996), Teoretyczne podstawy metody reprezentacyjnej.

1.3 Błędy

Na całkowity błąd badania składają się:

- błąd pomiaru i efekt respondenta
- błąd przetwarzania
- błąd pokrycia
- błąd losowania
- błąd braku odpowiedzi
- błąd dopasowania

1.3.1 Błąd pomiaru (measurement error)

Jest to różnica pomiędzy tym, co chcemy zmierzyć, a tym co otrzymujemy od respondenta. Np. w pytaniu: *Have you ever, even once, used any form of cocaine?* (National Survey on Drug Use and Health) raczej nie spodziewajamy się szczerych odpowiedzi.

Do tego dochodi response bias (efekt respondenta) czyli wpływ respondenta na otrzymywane wyniki. Przy powtarzalnych badaniach respondenci mogą nauczyć się kwestionariusza i odpowiadać tak, żeby skrócić wywiad.

1.3.2 Błąd przetwarzania (processing error)

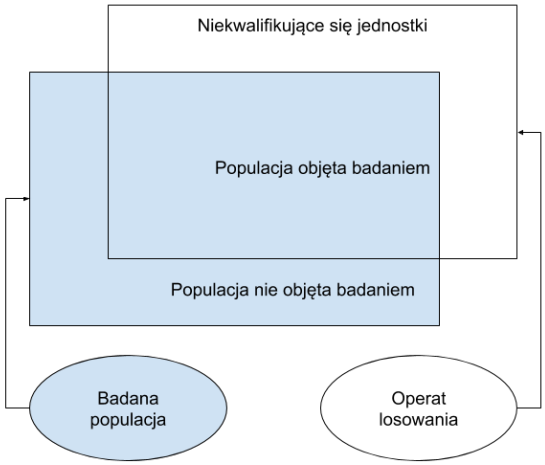
Jest to błąd wynikający z analizy odpowiedzi bez szerszego kontekstu oraz błąd kodowania pytań otwartych.

Jeśli podczas analizy kwestionariusza zobaczymy, że ktoś na pytanie *Ile razy dziennie doświadczasz agresji ze strony innych ludzi?* (National Crime Victimization Survey) odpowiedział, że 20 to będziemy mocno zaskoczeni. Nasze zaskoczenie zamieni się w zakłopotanie, kiedy zauważymy, że ta odpowiedź dotyczy respondenta płci męskiej. Na szczęście wykonywany zawód: ochroniarz w klubie nocnym pozwoli na zaakceptowanie takiego stanu rzeczy.

Inną kwestią są pytania otwarte: *Czy w ciągu ostatnich miesięcy słyszałeś o jakichkolwiek zmianach w nastrojach rynkowych?* (Surveys of Consumers). Z racji tego, że występuje trudność w analizie takich pytań często kategoryzuje się taką odpowiedź np. do trzech kategorii: pozytywna, neutralna, negatywna. Problem w tym, że tego przyporządkowania dokonuje pracownik firmy badawczej i często jest to jego subiektywny wybór.

1.3.3 Błąd pokrycia (coverage error)

Polega na objęciu badaniem niekompletnej zbiorowości. Jeżeli za operat losowania przyjęlibyśmy książkę telefoniczną to badaniem nie obejmiemy osób, które nie posiadają telefonu (populacja nie objęta badaniem). W książce telefonicznej znajdują się także numery do przedsiębiorstw, które będą niekwalifikującymi się jednostkami.



Coverage bias to różnica pomiędzy wynikami dla jednostek objętych i nieobjętych badaniem - w teorii do obliczenia.

1.3.4 Błąd losowania (sampling error)

Sampling bias występuje jeżli jednostki w operacie losowania nie mają szans na dostanie się do próby np. ze względu na zbyt małe prawdopodobieństwo wylosowania.

Sampling variance polega na tym, że każdorazowe losowanie próby będzie dawało odmienne wyniki.

1.3.5 Błąd braku odpowiedzi (nonresponse error)

Uczniowie nieobecni podczas testu kompetencji z matematyki mogą celowo opuszczać ten dzień w szkole ze względu na świadomość mniejszych umiejętności. Uzyskany przez szkołę wynik będzie z tego względu wyższy niż w przypadku obecności wszystkich uczniów.

1.3.6 Błąd dopasowania (adjustement error)

Wykorzystanie danych na temat populacji, wskaźnika kompletności w poprawie jakości próby może spowodować przeszacowanie lub niedoszacowanie wyników dla określonych grup.

Chapter 2

Wprowadzenie do R

Prezentacja

2.1 Wczytanie danych

Wykorzystamy zbiór danych dotyczących wyników drugiej tury wyborów prezydenckich w Polsce na poziomie obwodów wyborczych. Dane pochodzą z serwisu PKW, a plik Excel znajduje się tutaj.

Dane wczytujemy z wykorzystaniem pakietu *readxl*, a następnie czyścimy nazwy kolumn oraz wybieramy te zawierające kluczowe informacje:

```
library(tidyverse)

wp <- readxl::read_xlsx("data/wybory2020.xlsx") %>%
  janitor::clean_names(.) %>%
  select(symbol_kontrolny:percent_glosow_niewaznych, percent_glosow_waznych:rafal_kazimierz_trzas

summary(wp)
```

```
##  symbol_kontrolny       nr_okw       kod_teryt          typ_gminy
##  Length:27227       Min.   : 1.00   Length:27227       Length:27227
##  Class :character   1st Qu.:14.00   Class :character   Class :character
##  Mode  :character   Median :26.00   Mode  :character   Mode  :character
##                     Mean   :25.45
##                     3rd Qu.:37.00
##                     Max.   :49.00
##
##   numer_obwodu     typ_obszaru         typ_obwodu          siedziba
##  Min.   :   1.00   Length:27227       Length:27227       Length:27227
##  1st Qu.:   3.00   Class :character   Class :character   Class :character
```

```
##  Median :   7.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :  38.29
##  3rd Qu.:  18.00
##  Max.   :1147.00
##
##     gmina              powiat          wojewodztwo          frekwencja
##  Length:27227       Length:27227       Length:27227       Min.   :  0.00
##  Class :character   Class :character   Class :character   1st Qu.: 61.82
##  Mode  :character   Mode  :character   Mode  :character   Median : 67.18
##                                                           Mean   : 66.42
##                                                           3rd Qu.: 72.14
##                                                           Max.   :100.00
##
##  percent_glosow_niewaznych percent_glosow_waznych andrzej_sebastian_duda
##  Min.   : 0.0000           Min.   : 66.67         Min.   :  0.00
##  1st Qu.: 0.5300           1st Qu.: 98.87         1st Qu.: 42.59
##  Median : 0.8200           Median : 99.18         Median : 54.81
##  Mean   : 0.9358           Mean   : 99.06         Mean   : 56.64
##  3rd Qu.: 1.1300           3rd Qu.: 99.47         3rd Qu.: 70.85
##  Max.   :33.3300           Max.   :100.00         Max.   :100.00
##  NA's   :3                 NA's   :3              NA's   :3
##  rafal_kazimierz_trzaskowski
##  Min.   :  0.00
##  1st Qu.: 29.15
##  Median : 45.19
##  Mean   : 43.36
##  3rd Qu.: 57.41
##  Max.   :100.00
##  NA's   :3
```

Spróbujmy odpowiedzieć na kilka pytań dotyczących analizowanych danych.

1. Ile obwodów głosowania miało frekwencję powyżej 80%?

```
wp %>%
  filter(frekwencja > 80) %>%
  count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  1189
```

2. Ile obwodów głosowania znajduje się w Poznaniu?

```
wp %>%
  filter(powiat == "Poznań") %>%
  count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   251
```

3. Ile jest obwodów według typu obszaru?

```
wp %>%
  count(typ_obszaru)
```

```
## # A tibble: 6 x 2
##   typ_obszaru                     n
##   <chr>                       <int>
## 1 dzielnica w m.st. Warszawa    771
## 2 miasto                      12538
## 3 miasto i wieś                  82
## 4 statek                          8
## 5 wieś                        13661
## 6 zagranica                     167
```

4. Jaka była średnia frekwencja w województwach?

```
wp %>%
  group_by(wojewodztwo) %>%
  summarise(srednia_frekwencja=mean(frekwencja)) %>%
  arrange(srednia_frekwencja)
```

```
## # A tibble: 16 x 2
##    wojewodztwo         srednia_frekwencja
##    <chr>                            <dbl>
##  1 opolskie                          59.0
##  2 warmińsko-mazurskie               59.9
##  3 podlaskie                         63.2
##  4 lubuskie                          63.3
##  5 zachodniopomorskie                63.7
##  6 kujawsko-pomorskie                63.8
##  7 lubelskie                         64.8
##  8 dolnośląskie                      65.2
##  9 świętokrzyskie                    65.5
## 10 podkarpackie                      65.8
## 11 śląskie                           66.1
## 12 pomorskie                         67.0
## 13 wielkopolskie                     67.5
## 14 łódzkie                           68.7
## 15 małopolskie                       69.6
## 16 mazowieckie                       72.0
```

5. Gdzie była największa różnica pomiędzy kandydatami?

```
diff <- wp %>%
  mutate(roznica = abs(andrzej_sebastian_duda-rafal_kazimierz_trzaskowski))
```