

Estymacja wielkości populacji z wykorzystaniem pakietu uncounted

Zofia Danek | lic. Adrianna Mermela
dr Maciej Beręsewicz, prof. UEP

Ośrodek Metodologii Badań Ludnościowych
Urząd Statystyczny w Poznaniu

Agenda

Motywacja badawcza

Metody capture-recapture

Przykład zastosowania pakietu uncounged

- ▶ **Zmiana paradygmatu w statystyce publicznej**
 - ▶ Poszukiwanie alternatyw dla tradycyjnego spisu powszechnego
 - ▶ Wykorzystanie potencjału cyfryzacji administracji publicznej
- ▶ **Ograniczenia tradycyjnych metod**
 - ▶ Wysokie koszty realizacji
 - ▶ Braki odpowiedzi
- ▶ **Wyzwania migracyjne**
 - ▶ Wysoka mobilność populacji
 - ▶ Trudno dostępne grupy ludności (ang. *hard-to-reach populations*)

Metody capture-recapture

▶ Pochodzenie i zastosowania

- ▶ Petersen (1896) – populacje ryb
- ▶ Lincoln (1930) – populacje ptaków

▶ Idea

- ▶ Wielokrotne "łapanie" jednostek z populacji
- ▶ Identyfikacja jednostek obserwowanych wcześniej
- ▶ Szacowanie na podstawie powtórnych wyłapań oznaczonych jednostek

Naiwny estymator Lincolna-Petersena

$$\hat{N}_{LP} = \frac{n_{1+}n_{+1}}{n_{11}}$$

$$n_{1+} = n_{11} + n_{10}$$

$$n_{+1} = n_{11} + n_{01}$$

Lista 1 \ Lista 2	tak	nie
tak	n_{11}	n_{10}
nie	n_{01}	n_{00}

Tabela kontyngencji dla dwóch list

- ▶ **Możliwość wykorzystania różnych rejestrów**

- ▶ ZUS, PESEL, NFZ

- ▶ **Problemy**

- ▶ **Częściowy brak identyfikatorów** – trudności w łączeniu rekordów (np. brak identyfikatorów PESEL)
 - ▶ **Duplikaty i błędy w danych** – jedna osoba w wielu rekordach (np. w danych ZUS, raz z PESELelem a drugi raz bez)
 - ▶ **Nierówne prawdopodobieństwa** – różne szanse pojawienia się na listach

Błędy łączenia (ang. *linkage errors*)

▶ **Parametr α**

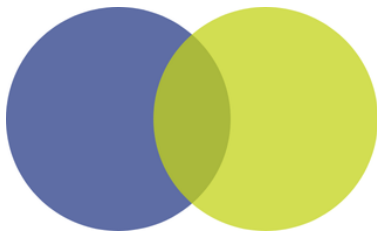
- ▶ Prawdopodobieństwo poprawnego zidentyfikowania prawdziwego dopasowania (ang. *true link*)
- ▶ $1 - \alpha$ - prawdopodobieństwo pominięcia prawdziwego dopasowania (ang. *false non-link*)

▶ **Parametr β**

- ▶ Prawdopodobieństwo błędnego połączenia niepasujących rekordów (ang. *false link*)

▶ **Dla trzech list:** wektory $[\alpha_1, \alpha_2]$ i $[\beta_1, \beta_2]$ dla każdej fazy łączenia

Klasyfikacja metod według liczby źródeł



DSE

Dual-System Estimation

Dwa źródła



MSE

Multiple System Estimation

Trzy (i więcej) źródeł

- ▶ **Naive Lincoln-Petersen** (Petersen 1896; Lincoln 1930)
 - ▶ Założenie doskonałego łączenia rekordów
 - ▶ Brak korekty błędów łączenia rekordów
- ▶ **Chapman** (Chapman 1951)
 - ▶ Korekta obciążenia dla małych próbek
 - ▶ Brak korekty błędów łączenia rekordów
- ▶ **Ding-Fienberg** (Ding i in. 1994)
 - ▶ Uwzględnienie fałszywych połączeń i braków łączenia rekordów (jednokierunkowych)
- ▶ **Modified Ding-Fienberg** (Di Consiglio i in. 2015)
 - ▶ Uwzględnienie fałszywych połączeń i braków łączenia rekordów (symetrycznych)
- ▶ **Asymmetric Correction** (Wolf i in. 2019)
 - ▶ Uwzględnienie fałszywych połączeń i braków łączenia rekordów (asymetrycznych)

- ▶ **Naive Lincoln-Petersen** (Fienberg 1972)
 - ▶ Założenie doskonałego łączenia rekordów
 - ▶ Brak korekty błędów łączenia rekordów
- ▶ **Chapman** (Wolf i in. 2019)
 - ▶ Korekta obciążenia dla małych próbek
- ▶ **Ding-Fienberg** (Fienberg i Ding 1996)
 - ▶ Uwzględnienie braków łączenia rekordów
- ▶ **Modified Ding-Fienberg** (Di Consiglio i in. 2018)
 - ▶ Uwzględnienie braków łączenia i fałszywych połączeń rekordów
- ▶ **Trivariate Bernoulli Models** (Chatterjee i in. 2019)
 - ▶ Uwzględnienie zależności między listami

Główne funkcje:

- ▶ `two_sources()`
 - ▶ 5 estymatorów: naive, Chapman, Ding-Fienberg, Modified DF, Asymmetric
- ▶ `three_sources()`
 - ▶ 4 estymatory: naive, Chapman, Ding-Fienberg, Modified DF
 - ▶ Modele log-liniowe z elastyczną specyfikacją interakcji
- ▶ `tbm()`
 - ▶ Modele TBM-1 i TBM-2 uwzględniające zależności między listami

Dodatkowe funkcjonalności:

- ▶ `generate_data()` – symulacje z kontrolowanymi błędami łączenia
- ▶ Analityczne i bootstrapowe obliczanie wariancji
- ▶ Błędy standardowe i przedziały ufności

Przykład: 1990 U.S. Census, St. Louis (Di Consiglio i in. 2018)

Dual System Estimation

Method: Modified Ding-Fienberg estimator (symmetric two-way correction)

Observed Data:

Counts: $n_{11} = 487$, $n_{10} = 129$, $n_{01} = 217$

Total: 833

Population Estimates:

$N_{\text{hat}} = 898$ (SE = 11)

95% CI for N: (877, 919)

$n_{00_hat} = 65$

Capture Probabilities:

$p_1 = 0.69$

$p_2 = 0.78$

Model Specifications:

Linkage errors: $\alpha = 0.9966$, $\beta = 0.0161$

Variance method: analytic

Symulacja – weryfikacja estymatorów DSE

- ▶ 10,000 zbiorów danych z błędami jednokierunkowymi (OC)
- ▶ $N = 1000$, $p = (0.7, 0.8)$, $\alpha = 0.95$, $\beta = 0.02$

Estymator	$\bar{N}$	$\overline{SE}$
Naive (LP)	1047.4	12.7
Chapman	1047.3	12.7
Ding-Fienberg	1000.3	10.4
Modified DF	1009.8	10.8
Asymmetric	1009.8	10.8

Dziękujemy za uwagę

<https://poznan.stat.gov.pl/osrodki/osrodek-metodologii-badan-ludnosciowych/>

Bibliografia I

- Chapman, D. G. (1951). *Some Properties of the Hypergeometric Distribution with Applications to Zoological Sample Censuses*. Berkeley: University of California Press.
- Chatterjee, Kiranmoy i Prajamitra Bhuyan (paż. 2019). "On the Estimation of Population Size from a Dependent Triple-Record System". W: *Journal of the Royal Statistical Society Series A: Statistics in Society* 182.4, s. 1487–1501. DOI: 10.1111/rssa.12472.
- Di Consiglio, L. i T. Tuoto (2015). "Coverage Evaluation on Probabilistically Linked Data". W: *Journal of Official Statistics* 31.3, s. 415–429.
- (2018). "Population size estimation and linkage errors: the multiple lists case". W: *Journal of Official Statistics* 34.4, s. 889–908.
- Ding, Y. i S. E. Fienberg (1994). "Dual system estimation of Census undercount in the presence of matching error". W: *Survey Methodology* 20, s. 149–158.
- Fienberg, S. E. (1972). "The Multiple Recapture Census for Closed Populations and Incomplete 2^k Contingency Tables". W: *Biometrika* 59.3, s. 591–603. DOI: 10.1093/biomet/59.3.591.

Bibliografia II

- Fienberg, S. E. i Y. Ding (1996). "Multiple sample estimation of population and census undercount in the presence of matching error". W: *Proceedings of 1994 Annual research conference and CASIC technologies Interchange*. Bureau of Census, US. Lincoln, F. C. (1930). *Calculating waterfowl abundance on the basis of banding returns*. Spraw. tech. 118. US Department of Agriculture.
- Petersen, C. G. J. (1896). "The yearly immigration of young plaice in the Limfjord from the German sea". W: *Rept. Danish Biol. Sta.* 6, s. 1–48.
- Wolf, P.-P. de, J. van der Laan i D. Zult (2019). "Connecting Correction Methods for Linkage Error in Capture-Recapture". W: *Journal of Official Statistics* 35.3, s. 577–597. DOI: 10.2478/jos-2019-0024.