mgr Wiesława Dąbała
Centrum Badania Opinii Społecznej
Warszawa
E-mail: w.dabala@cbos.pl

Metody imputacji brakujących danych w próbach z niepełną realizacją. Wnioski i pytania

Niech:

X_{ij} — wartość w j-tej zmiennej dla i-tej jednostki wylosowanej do próby, $X_{ij} = 0$ — oznaczenie braków danych;

i — numer jednostki, $i = 1, 2, \ldots, n$;

n — liczebność próby wylosowanej;

j — numer zmiennej badania, $j = 1, 2, \ldots, k$;

k — liczba zmiennych w badaniu.

Wyodrębniane są dwa rodzaje braków danych. Są to:

— braki danych we wszystkich zmiennych dla jednostki populacji wylosowanej do próby, od której nie uzyskano żadnych informacji. Wówczas dla ustalonego i, dla każdego $j = 1, 2, \ldots, k$, wartości X_{ij} mają wartość zero. W takiej sytuacji liczebność próby zrealizowanej jest mniejsza od liczebności próby wylosowanej.

— braki danych tylko w części zmiennych, dla jednostek populacji wylosowanych do próby, które objęto badaniem. Wówczas dla ustalonego i tylko dla niektórych zmiennych j mamy $X_{ij} = 0$. Podstawienia danych oszacowanych w miejsce danych brakujących stosowane są na ogół dla niewielkiej liczby przypadków dla jednostki lub dla prób z wysokim wskaźnikiem realizacji zamiast ważenia danych.

W zależności od charakteru braków danych (systematycznego, losowego itp.) stosowane są różne metody oszacowań danych brakujących. Są to zarówno metody statystycznej analizy danych jak i podstawienia losowo wybranych danych lub jednostek, z grup jednostek o licznych cechach podobieństwa. Wykorzystywane są także dane zewnętrzne.

Zaletą stosowania podstawień jest możliwość przejrzystej prezentacji wyników, co jest dobrze postrzegane przez odbiorców wyników badań marketingowych, szczególnie w przypadkach badań na próbach o niewielkich liczebnościach ze zbiorowości o złożonych strukturach lub przy wykorzystaniu złożonych metod analiz.

Zwiększanie się liczby braków danych powoduje występowanie różnic oszacowań z zastosowaniem różnych metod.

Oprócz zastrzeżeń natury etycznej powstaje też wiele wątpliwości i pytań o właściwy dobór metod, ich efektywność, występujące ograniczenia stosowania, możliwości oszacowania błędów występujących w wyniku stosowania metod imputacji.

Literatura

[1] A. Balicki, *Metody imputacji brakujących danych w badaniach statystycznych*, Wiadomości Statystyczne nr. 9, wrzesień 2004 r. GUS, PTS.

[2] Cz. Bracha, *Metoda reprezentacyjna w badaniach marketingowych i społecznych*, Efekt 1998.

[3] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning. Data Mining, Inference and Prediction*, Springer, New York–Berlin–Heidelberg 2001.