Algorytmiczne Zastosowania Łańcuchów Markowa
Projekt 10 - GUS

W badaniach przeprowadzanych przez Główny Urząd Statystyczny, często ma się do czynienia z sytuacją, gdy badana populacja (której jakaś cecha nas interesuje) posiada warstwy. Przedsiębiorstwa są podzielone na branże, państwo na województwa itp. Konstruuje się estymator interesującej nas cechy który z definicji jest nieobciążony. W takiej sytuacji poszukuje się estymatora, którego wariancja będzie najmniejsza.

Część 1 W przypadku gdy w badania z losowaniem jednostopniowym problem sprowadza się do minimalizacji wyrażenia (patrz praca J. Wesołowski, R. Wieczorkowski, W. Wójciak *Recursive Neyman algorithm for optimum sample allocation under box constraints on sample sizes in strata*)

$$\sum_{h=1}^{H} \frac{A_h^2}{x_h}$$

gdzie H to liczba warstw, A_h są dane (współczynniki wariancji w warstwach), a x_h to liczba elementów z danej warstwy wybrana do badania. Celem jest minimalizacja powyższej wariancji przy warunku $\sum_{h \in H} x_h = n$, czyli ustalonej liczności próby. Długo nierozwiązany był problem optymalizacji przy nadanych warunkach na liczności w warstwach $m_h \leq x_h \leq M_h$. Wyżej wymieniona praca rozwiązuje ten problem i jest zaimplementowana jako algorytm RNABOX w pakiecie *stratallo* dostępnym w CRAN.

Cel 1: Porównać dzianie algorytmu simulated annealing z algorytmem RNABOX, w szczególności porównać szybkość i dokładność rozwiązania.

Część 2 W przypadku losowania dwustopniowego z warstwami na pierwszym stopniu (na przykład losujemy szkoły, gdzie warstwami są województwa, a następnie ze szkół losujemy uczniów) wariancja ma bardziej skomplikowaną postać:

$$\sum_{h=1}^{H} \left(\frac{1}{m_h} - \frac{1}{M_h} \right) M_h^2 D_h^2 + \sum_{h=1}^{H} \frac{M_h}{m_h} \sum_{j=1}^{M_h} \left(\frac{1}{n_{h,j}} - \frac{1}{N_{h,j}} \right) N_{h,j}^2 S_{h,j}^2$$

optymalna alokacja znana jest tylko w przypadku bez ograniczeń na liczebności w warstwach. W powyższym wzorze dla przykładu ze szkołami:

- M_h to znana liczba szkół w województwie numer h oraz D_h^2 to znany współczynnik wariancji dla szkół w tym województwie.

- $N_{h,j}$ to znana liczba uczniów w j-tej szkole w h-tym województwie oraz $S_{h,j}$ to znany współczynnik wariancji dla uczniów tej szkoły.
- Naszym zadaniem jest minimalizacja powyższej wariancji (jako funkcji $m_1, \ldots, m_h$ i $n_{1,1}, \ldots, n_{h,1}, \ldots$) przy warunku

  $\sum_{h=1}^{H} \frac{m_h}{M_h} \sum_{j=1}^{M_h} N_{h,j} = n_I$ (oczekiwana liczba uczniów w jednostkach pierwszego stopnia)

  $\sum_{h=1}^{H} \frac{m_h}{M_h} \sum_{j=1}^{M_h} n_{h,j} = n$ (oczekiwana liczba uczniów w ostatecznej próbce).

  Opis badania (patrz W. Niemiro, J. Wesołowski *FIXED PRECISION OPTIMAL ALLOCATION IN TWO-STAGE SAMPLING*, uwaga: ta praca dotyczy innego, związanego zagadnienia, w którym interesują nas średnie w województwach, i zachowanie porównywalności wariancji estymatorów w poszczególnych województwach, ale powyższy problem jest opisany we wstępie pracy). Algorytm rozwiązujący powyższy problem jest znany tylko bez uwzględniania naturalnych ograniczeń górnych $m_h \leq M_h$ oraz $n_{h,j} \leq N_{h,j}$.

  Cel 2: Zaprojektować i zaimplementować algorytm symulowanego wyżarzania dla powyższego problemu z ograniczeniami górnymi.

  Uwaga: pierwszy warunek

  $$\sum_{h=1}^{H} \frac{m_h}{M_h} \sum_{j=1}^{M_h} N_{h,j} = n_I$$

  można zastąpić warunkiem $\sum_{h=1}^{H} m_h = m$, czyli ustalamy liczbę szkół w próbce.