

Prof. dr hab. Wojciech Zieliński
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
Katedra Ekonometrii i Statystyki
02-767 Warszawa, Nowoursynowska 166

RECENZJA ROZPRAWY DOKTORSKIEJ
MGR. WOJCIECHA WÓJCIAKA
*Multi-Domain Optimum Sample Allocation with Controlled Precision
under Upper-Bound Constrains*
Warsaw University of Technology

1. Podstawa sporządzenia recenzji

Podstawą sporządzenia recenzji jest uchwała Rady Naukowej Dyscypliny Matematyka Politechniki Warszawskiej nr 43/III/2026 z dnia 23 stycznia 2026 roku. W swoim piśmie Przewodniczący Rady Naukowej zwraca się z prośbą o sporządzenie recenzji rozprawy doktorskiej mgra Wojciech Wójciaka pt. *Multi-Domain Optimum Sample Allocation with Controlled Precision under Upper-Bound Constrains*

Postępowanie prowadzone jest w dyscyplinie matematyka zgodnie z ustawą Prawo o Szkolnictwie Wyższym i Nauce z dnia 20 lipca 2018 r. (Dz.U z 2023 r., poz. 742 z późn. zm.). Zgodnie z artykułem 187 Ustawy opinia dotycząca danej rozprawy doktorskiej powinna zawierać następujące elementy:

- czy rozprawa doktorska prezentuje ogólną wiedzę teoretyczną kandydata w dyscyplinie matematyka;
- czy rozprawa doktorska wykazuje umiejętność samodzielnego prowadzenia pracy naukowej przez osobę ubiegającą się o nadanie stopnia doktora;
- czy rozprawa doktorska stanowi oryginalne rozwiązanie problemu naukowego.

2. Struktura rozprawy doktorskiej oraz jej ogólna ocena

Recenzowana rozprawa doktorska składa się ze siedmiu rozdziałów, siedmiu dodatków i bibliografii.

W rozdziale 1. „Introduction” sformułowany jest główny cel pracy i problem badawczy. W tym rozdziale wprowadzone są niezbędne pojęcia i oznaczenia. Przedstawiony jest także przegląd odpowiedniej literatury.

W rozdziale 2. „Population — Total Sample Size Model” wprowadzony jest rozważany w pracy model populacji. Analizowana jest skończona N elementowa populacja podzielona na D domen, które składają się z H warstw. Liczebności poszczególnych części populacji (tzn. domen i warstw) traktowane są jako znane. Liczba warstw w każdej domenie może być inna. W populacji obserwowana pewna cecha ilościowa, której odchylenia standardowe w poszczególnych warstwach jest znane.

Rozdział 3. „The Controlled-Precision Domain Allocation Problem” zawiera najważniejsze wyniki dotyczące optymalnej alokacji próby. Ten rozdział jest kluczowy dla całej dysertacji.

Rozdziały 4.-6. poświęcone są problemom numerycznej implementacji rozwiązań uzyskanych we wcześniejszych rozdziałach. W tych rozdziałach opisany jest algorytm numerycznego poszukiwania optymalnej alokacji próby, zbadane są jego własności oraz zanalizowana poprawność jego działania.

Rozdział 7. stanowi podsumowanie pracy oraz przedstawia dalsze kierunki badań nad zagadnieniem optymalnej alokacji próby.

Rozprawę doktorską kończy Bibliografia zawierająca 53 pozycje literaturowe. Autor wykazuje trzy swoje publikacje współautorskie oraz jedną publikację samodzielną.

Pierwsze trzy rozdziały mają charakter matematyczny, kolejne trzy są rozdziałami o charakterze informatycznym.

Rozprawa napisana jest w języku angielskim. Układ pracy jest bardzo czytelny i klarowny: podział na rozdziały i paragrafy jest jak najbardziej prawidłowy. Praca napisana jest w LATEXu co szalenie zwiększa jej przejrzystość i czytelność. Należy zwrócić uwagę, że przedstawiona do oceny praca wydana jest w przyjemnej dla Czytelnika formie książkowej.

3. Podstawowe wyniki

Praca poświęcona jest zagadnieniom typowym dla takiego działu statystyki matematycznej jakim jest metoda reprezentacyjna, czyli statystyka populacji skończonych. Jest to dział intensywnie eksploatowany przede wszystkim w badaniach ekonomicznych oraz socjologicznych. Ogólnie mówiąc, w metodzie reprezentacyjnej dokonuje

się wyboru podzbioru populacji, na podstawie którego prowadzone jest wnioskowanie o całej zbiorowości.

Bardziej formalnie, w skończonej N elementowej populacji U badacza interesuje pewna cecha ilościowa y o dodatnich wartościach. Wielkość populacji N jest znana. Celem badania jest oszacowanie wartości globalnej t tej cechy w całej populacji. Badanie prowadzone jest metodą reprezentacyjną, tzn. z populacji pobierana jest próba losowa i na podstawie zaobserwowanych w próbie wartości cechy szacowana jest poszukiwana wartość. W takich badaniach wyrywkowych eksperymentatorowi zależy na jak najdokładniejszej estymacji, tzn. wynik estymacji ma być jak najbliższy rzeczywistej szacowanej wielkości. Dokładność estymacji zależy głównej mierze od wyboru estymatora, wielkości próby i wskaźnika dokładności estymatora (typowo przyjmuje się jego wariancję, odchylenie standardowe lub współczynnik zmienności). Jako podstawowy miernik „dobroci” estymacji przyjęto w pracy współczynnik zmienności estymatora.

W pracy jako estymator wartości globalnej t przyjęto estymator Horvitz-Thompsona, a także ustaloną wielkość próby n . Estymator ten jest w każdym schemacie losowania estymatorem nieobciążonym, co pozwala jako miarę dokładności estymatora przyjąć klasycznie jego wariancję, a nie błąd średniokwadratowy. W związku z tym w celu poprawy jakości estymacji wykorzystywany jest podział populacji na „w miarę” jednorodnych warstw. To czym może sterować eksperymentator w celu uzyskania dokładnego oszacowania estymowanej wartości globalnej jest odpowiednia alokacja n obiektów próby pomiędzy domeną i warstwami. Pierwsze wyniki w tym kierunku uzyskali już Chuprov (1923) i niezależnie Neyman (1934). Po tych pionierskich pracach uzyskano wiele wyników dotyczących różnych populacji, różnych cech czy różnych kryteriów estymacji.

Bardziej formalnie, w pracy rozważana jest skończona populacja złożona z N obiektów, w której interesuje nas pewna cecha y . Zadanie polega na jak najdokładniejszej estymacji nieznanej wartości globalnej tej cechy w całej populacji na podstawie n elementowej próby. Populacja podzielona jest na D domen $\{d = 1, \dots, D\}$. Ponadto, każda domena jest podzielona na H_d $\{h = 1, \dots, H_d\}$ warstw. Liczności domen oraz warstw traktujemy jako znane: $\sum_{d=1}^D \sum_{h=1}^{H_d} N_{d,h} = N$. W każdej warstwie stosujemy schemat losowania prostego bez zwracania. W ten sposób mamy $\sum_{d=1}^D H_d$ niezależnych prób o liczebnościach $n_{d,h}$, które posłużą do estymacji wartości globalnej t . Jako estymator wartości globalnej t przyjmujemy estymator Horvitz-Thompsona:

$$\hat{t} = \sum_{d=1}^D \sum_{h=1}^{H_d} \frac{N_{d,h}}{n_{d,h}} \sum_{k=1}^{n_{d,h}} y_k.$$

Wariancja tego estymatora wynosi

$$\text{Var}(\hat{t}) = \sum_{d=1}^D \sum_{h=1}^{H_d} \left(\frac{1}{n_{d,h}} - \frac{1}{N_{d,h}} \right) N_{d,h}^2 S_{d,h}^2,$$

gdzie $S_{d,h}^2$ jest wariancją badanej cechy w d -tej domenie i h -tej warstwie tej domeny.

Ponieważ wariancja jest bezwzględną miarą zróżnicowania cechy co może w pewnych zagadnieniach praktycznych nie być miarą adekwatną w analizach porównawczych, więc stosowany jest współczynnik zmienności. Ten ostatni współczynnik jest miarą względną, która pozwala porównywać zróżnicowanie cech w różnych jednostkach lub o odmiennych rzędach wielkości. Dlatego w pracy zdecydowano się na wykorzystanie właśnie tego miernika jako miary zróżnicowania. W związku z tym, dla każdej domeny określamy kwadrat współczynnika zmienności

$$V_d = \frac{\text{Var}(\hat{t}_d)}{(E[\hat{t}_d])^2}$$

i zakładamy, że $V_d = \kappa_d T$ ($d = 1, \dots, D$). Tutaj κ_d są znanymi nieujemnymi stałymi sumującymi się do jedynki (wagami), a T jest pewnym nieujemnym skalar. Liczbę T można sobie wyobrażać jako „ogólny” współczynnik zmienności w całej populacji, natomiast V_d udziały poszczególnych domen w tej ogólnej zmienności.

Autor formułuje problem badawczy w następujący sposób.

Problem 1.1.1. Znaleźć licznosci prób $\{n_{d,h}, h = 1, \dots, H_d, d = 1, \dots, D\}$ oraz skalar T takie, że

$$\begin{aligned} T &= \min! \\ \sum_{d=1}^D \sum_{h=1}^{H_d} n_{d,h} &= n, \\ \frac{1}{t_d^2 \kappa_d} \sum_{h=1}^{H_d} \left(\frac{1}{n_{d,h}} - \frac{1}{N_{d,h}} \right) N_{d,h}^2 S_{d,h}^2 &= T, \quad d = 1, \dots, D \\ 0 < n_{d,h} &\leq N_{d,h}, \quad h = 1, \dots, H_d, \quad d = 1, \dots, D \end{aligned}$$

Autor w pracy rozważa nieco ogólniejsze zagadnienie. Jako podstawowy miernik do broci rozwiązania przyjmuje współczynnik zmienności estymatora oraz dopuszcza by licznosci były dowolnymi liczbami rzeczywistymi. To drugie znacznie ułatwia całą matematykę. Czyli analizowane zagadnienie jest sformułowane w następujący sposób. Zastosowałem tutaj zamiast zapisu proponowanego przez Autora zapis równoważny częściej spotykany w literaturze przedmiotu.

Problem 3.1.1. Znaleźć nieujemne liczby $\{x_{d,h}, h = 1, \dots, H_d, d = 1, \dots, D\}$ oraz skalar T takie, że

$$\begin{aligned} T &= \min! \\ \sum_{d=1}^D \sum_{h=1}^{H_d} x_{d,h} - n &= 0, \\ \sum_{h=1}^{H_d} \left(\frac{1}{x_{d,h}} - \frac{1}{N_{d,h}} \right) \frac{|N_{d,h} S_{d,h}|^2}{t_d^2 \kappa_d} - T &= 0, \quad d = 1, \dots, D \\ x_{d,h} - N_{d,h} &\leq 0, \quad h = 1, \dots, H_d, \quad d = 1, \dots, D \end{aligned}$$

Uzyskane liczby x mają interpretację liczności prób z poszczególnych domen i warstw.

W rozdziałach 2 i 3 przedstawiona jest szczegółowa teoretyczna analiza problemu (przede wszystkim) 3.1.1. Zagadnienie to nazwane zostało jako *CPDA: The Controlled-Precision Domain Allocation Problem*. Doktorant przeprowadził bardzo dokładną analizę tego zagadnienia. Ważne jest następujące stwierdzenie:

Propozycja 3.1.5. Dla rozważanego zagadnienia istnieje rozwiązanie problemu 3.1.1.

Czyli w rozważanym zagadnieniu optymalna alokacja próby istnieje. W pracy podane jest również analityczne rozwiązanie tego problemu. Dokładniej, zaproponowane jest pewne rozwiązanie, którego optymalność jest dowodzona w twierdzeniu 3.2.8.

Doktorant stworzył aplikację w Języku R pozwalającą na numeryczne rozwiązanie konkretnego zagadnienia. Pokazane jest działanie tej aplikacji na przykładzie populacji złożonej z dwóch domen, z których każda podzielona jest na dwie warstwy oraz na przykładzie, w którym są trzy domeny po dwie warstwy. Autor dokładnie zbadał poprawność zastosowanego algorytmu oraz działania programu. Stworzony przez Doktoranta program został dołączony do *Comprehensive R Archive Network* pod nazwą **stratallo**.

3. Ogólna ocena pracy

Pan Wojciech Wójciak podjął się rozwiązania zagadnienia bardzo ważnego dla praktyki. W badaniach empirycznych fundamentalnym pytaniem jest pytanie o realizację jak najdokładniejszej (precyzyjniejszej) estymacji wartości globalnej pewnej cechy ilościowej. Ponieważ wnioskowanie oparte jest na badaniach wrywkowych, więc powstaje problem zbierania danych. Jednym ze sposobów „dobrego” próbkowania jest podział populacji na jednorodną, mniejszą część. To realizowane jest przez podział

populacji na tak zwane domeny, a te domeny na jeszcze mniejsze grupy zwane warstwami. W zastosowaniach metody reprezentacyjnej pod pojęciem domeny rozumiane są na przykład województwa lub powiaty, zaś jako warstwy traktuje się na przykład podział na płeć, grupy wiekowe czy wykształcenie. Z reguły badana cecha jest „bardziej wyrównana” w domenach czy warstwach co pozwala na lepszą estymację wartości globalnej. W praktyce rodzi się pytanie jak zbierać dane z uwzględnieniem tego podziału, by uzyskać jak najbardziej dokładną ocenę nieznaną globalnej wartości badanej cechy w całej badanej populacji.

Doktorant z powodzeniem teoretycznie rozwiązuje postawione zadanie. Ponadto Doktorant przygotował komputerowy program pozwalający numerycznie rozwiązać zagadnienie.

4. Uwagi szczegółowe

Cała rozprawa napisana jest po angielsku ze zwróceniem szczególnej uwagi na poprawność językową. Do strony edytorskiej też nie można mieć zastrzeżeń. Jedyne co nasunęło mi się przy lekturze pracy doktorskiej Pana Wojciecha Wójciaka to komentarze sformułowane poniżej. Te uwagi mają raczej charakter uzupełniający do całej rozprawy i w niczym nie umniejszają wartości całej pracy.

Autor stosuje bardzo formalny język matematyczny. Taki język jest w miarę łatwo zrozumiały dla matematyka, ale niestety dla praktyka jest on mało przyswajalny i trochę zniechęcający. Gdyby Autor rozważał udostępnianie swoich (ważnych i cennych) wyników praktykom zajmującym się rozważanym zagadnieniem, to sugerowałbym zwiększenie czytelności poprzez pewne „odmatematyzowanie” języka.

Autor rozwiązuje zadanie, w którym przyjęto, że najlepsza dokładność estymacji tzn. wariancja (współczynnik zmienności) estymatora HT, jest minimalna, ale taka sama dla każdej domeny. Jest to ujęte m.in. w warunku 3.2b problemu 3.1.1. Innymi słowami rozwiązywane jest zadanie jednakowej precyzji estymacji we wszystkich domenach. To może być trudne do spełnienia, szczególnie gdy domeny są mocno zróżnicowane między sobą. Czy może lepsze rezultaty nie mogły być osiągnięte gdyby jednak dopuścić w każdej domenie inną maksymalną precyzję?

W każdej domenie i w każdej warstwie zastosowany jest schemat losowanie prostego bez zwracania. Brakuje jakiegoś komentarza na temat ewentualnego zastosowania innego schematu losowania.

Jednym z ważnych aspektów prowadzenia badań metodą reprezentacyjną jest wzięcie pod uwagę kosztów takich badań, a także budżetu przeznaczanego na te badania. Wydaje się, że wzięcie pod uwagę tego elementu przy planowaniu eksperymentu mogłoby być ciekawym rozszerzeniem uzyskanych wyników.

Rozwiązaniem problemu mogą być ułamkowe „liczności” prób, patrz na przykład punkt 4.2.1. na stronie 59. Co doradzić praktykowi w takiej sytuacji: jakieś zaokrąglenie? randomizację? czy jeszcze jakieś inne wybrnięcie z kłopotu?

W pracy Doktorant zajmuje się cechami ilościowymi takimi jak na przykład dochody, emerytury, plony itp. Ciekawym byłoby zastosowanie prowadzonych przez Autora rozważań do takich zagadnień jak np. popularność różnych osobistości czy oglądalności programów telewizyjnych. Pytanie jakie się rodzi to czy można w miarę łatwo zaadaptować uzyskane wyniki do takich zmiennych? Ze znaczeniem takiego problemu mieliśmy do czynienia chociażby w ostatnich wyborach prezydenckich.

5. Wnioski końcowe

Recenzowana rozprawa doktorska prezentuje ogólną wiedzę teoretyczną Kandydata w dyscyplinie matematyka z zakresu statystyki matematycznej. Świadczy o tym umiejętność ścisłego i formalnego formułowania problemu oraz wykorzystanie aparatu matematycznego niezbędnego do rozwiązania zadania. Ponadto Kandydat wykazał się umiejętnością zaawansowanego programowania w języku R.

Reasumując stwierdzam jednoznacznie, że dysertacja mgra Wojciecha Wójciaka *Multi-Domain Optimum Sample Allocation with Controlled Precision under Upper-Bound Constrains* spełnia wszystkie wymagania stawiane pracom doktorskim określone w art. 187 ustawy z dnia 20 lipca 2018 r. Prawo o Szkolnictwie Wyższym i Nauce (Dz.U. z 2023 r., poz. 742 z późn. zm.). Wnoszę zatem o dopuszczenie Pana Wojciecha Wójciaka do dalszych etapów przewodu doktorskiego.

6. Literatura

Pozycje wykorzystane w recenzji. Pozycji tych nie ma w spisie literatury załączonym przez Doktoranta.

- Chuprov, Aleksandr Aleksandrowicz (1923). “On the mathematical expectation of the moments of frequency distributions in the case of correlated observations”. *Metron* 2, s. 461–493, 646– 683
- Neyman, Jerzy (1934). “On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection”. *Journal of the Royal Statistical Society* 97.4, s. 558–625.

