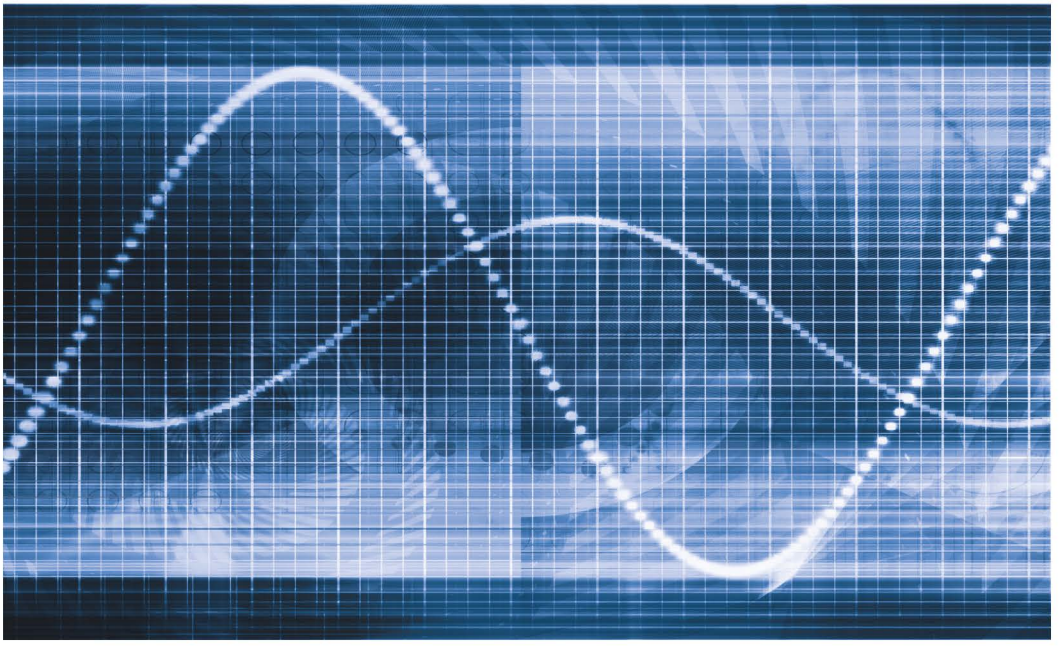
Ekonomia

Parametr wygładzania w estymacji jądrowej funkcji gęstości dla zmiennych losowych w badaniach ekonomicznych

Aleksandra Baszczyńska



WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

Parametr wygładzania w estymacji jądrowej funkcji gęstości dla zmiennych losowych w badaniach ekonomicznych

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

Ekonomia

Parametr wygładzania w estymacji jądrowej funkcji gęstości dla zmiennych losowych w badaniach ekonomicznych

Aleksandra Baszczyńska

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO
ŁÓDŹ 2016

Aleksandra Baszczyńska – Uniwersytet Łódzki, Wydział Ekonomiczno-Socjologiczny
Katedra Metod Statystycznych, 90-255 Łódź, ul. POW 3/5

RECENZENT
Grzegorz Kończak

REDAKTOR INICJUJĄCY
Monika Borowczyk

OPRACOWANIE REDAKCYJNE
Małgorzata Szymańska

SKŁAD I ŁAMANIE
Munda – Maciej Torz

PROJEKT OKŁADKI
Stämpfli Polska Sp. z o.o.
Zdjęcie wykorzystane na okładce: © Shutterstock.com

Wydane przez Wydawnictwo Uniwersytetu Łódzkiego
Wydanie I. W.07511.16.0.M

Ark. 7,2; ark. druk. 12,5

ISBN 978-83-8088-279-9
e-ISBN 978-83-8088-280-5

Wydawnictwo Uniwersytetu Łódzkiego
90-131 Łódź, ul. Lindleya 8
www.wydawnictwo.uni.lodz.pl
e-mail: ksiegarnia@uni.lodz.pl
tel. (42) 665 58 63

Spis treści

Indeks oznaczeń i symboli

$X_1, X_2, \ldots, X_n$	próba losowa
$x_1, x_2, \ldots, x_n$	realizacja próby losowej
$f = f(x)$	funkcja gęstości
$F(x)$	dystrybuanta
$\hat{F}(x)$	dystrybuanta empiryczna
$I_A(x)$	funkcja charakterystyczna zbioru A: $I_A(x) = \begin{cases} 1 & \text{gdy} \quad x \in A \\ 0 & \text{gdy} \quad x \notin A \end{cases}$
$\hat{f}(x)$	estymator funkcji gęstości
$\hat{f}_H(x)$	histogram
h_H	szerokość klasy w histogramie
$\hat{f}_n(x) = \hat{f}(x)$	estymator jądrowy funkcji gęstości
$\hat{f}_n^{(v)}(x)$	estymator jądrowy ν-tej pochodnej funkcji gęstości
$K = K(u)$	funkcja jądra
$h = h(n)$	parametr wygładzania
$S_{v,k}^{\mu}$	klasa funkcji jądra o rzędzie k i gładkości μ
$E[\hat{f}(x)]$	wartość oczekiwana estymatora funkcji gęstości
$B[\hat{f}(x)]$	obciążenie estymatora funkcji gęstości
$D^2[\hat{f}(x)]$	wariancja estymatora funkcji gęstości
$\|\|\mathbf{x}\|\|_p$	p-ta norma wektora $\mathbf{x}$: $\|\mathbf{x}\|_p = \left(\|x_1\|^p + \|x_2\|^p + \ldots + \|x_n\|^p\right)^{\frac{1}{p}}$
$\hat{f}_{Ri}(x)$	estymator jądrowy funkcji gęstości z odbiciem, $i = L, P$

$\hat{f}_B(x)$	estymator jądrowy funkcji gęstości z brzegową funkcją jądra
AE$[\hat{f}(x)]$	błąd bezwzględny estymatora funkcji gęstości
$MAE[\hat{f}(x)]$	średni błąd bezwzględny estymatora funkcji gęstości
$IAE[\hat{f}(x)]$	scałkowany błąd bezwzględny estymatora funkcji gęstości
$MIAE[\hat{f}(x)]$	scałkowany średni błąd bezwzględny estymatora funkcji gęstości
$SE[\hat{f}(x)]$	błąd kwadratowy estymatora funkcji gęstości
$MSE[\hat{f}(x)]$	błąd średniokwadratowy estymatora funkcji gęstości
$ISE[\hat{f}(x)]$	scałkowany błąd kwadratowy estymatora funkcji gęstości
$MISE[\hat{f}(x)]$	scałkowany błąd średniokwadratowy estymatora funkcji gęstości
$\kappa = R(K)$	scałkowany kwadrat funkcji jądra: $\kappa = \int_{-\infty}^{+\infty} K^2(u)du$
κ_k	k-ty moment zwykły funkcji jądra: $\kappa_k = \int_{-\infty}^{+\infty} u^k K(u)du$
$R(f^{(k)})$	scałkowany kwadrat k-tej pochodnej funkcji gęstości: $R(f^{(k)}) = \int_{-\infty}^{+\infty} \left[f^{(k)}(x)\right]^2 dx$
$K_J(u)$	jednostajna funkcja jądra
$K_T(u)$	trójkątna funkcja jądra
$K_E(u)$	funkcja jądra Epanecznikowa
$K_{DW}(u)$	dwuwagowa funkcja jądra
$K_{TW}(u)$	trójwagowa funkcja jądra
$K_G(u)$	gaussowska funkcja jądra
$K_C(u)$	kosinusowa funkcja jądra
$K_{ST}(u)$	funkcja jądra stopnia trzeciego
$K * K(u)$	splot funkcji jądra: $K * K(u) = K^{(2)}(u) = \int_{-\infty}^{+\infty} K(u-v)K(v)dv$
$n!!$	silnia podwójna: $n!! = \begin{cases} 1 \quad \text{dla} \quad n=0 \quad \text{lub} \quad n=1 \\ n(n-2)!! \quad \text{dla} \quad n \geq 2 \end{cases}$

$(a)_n$ symbol Pochhammera: $(a)_n = \prod_{j=0}^{n-1}(a+j)$

$\Gamma(z)$ funkcja gamma Eulera

$B(x, y)$ funkcja beta Eulera

$C_m^l(u)$ wielomian Gegenbauera:

$$C_m^l(u) = \frac{1}{\Gamma(l)} \sum_{i=0}^{\left[\frac{m}{2}\right]} \frac{(-1)^i \Gamma(l+m-i)(2u)^{m-2i}}{i!(m-2i)!}$$

$P_k(x)$ wielomian Legendre'a stopnia k : $P_k(x) = \sum_{r=0}^{k} p_r^k u^r$

$\phi(u)$ funkcja gęstości rozkładu normalnego standaryzowanego

h_{ISE} parametr wygładzania minimalizujący *ISE*

h_{MISE} parametr wygładzania minimalizujący *MISE*

h_{AMISE} parametr wygładzania minimalizujący *AMISE*

$h_{RR,\cdot}$ parametr wygładzania wyznaczony metodą odwołania do standardowego rozkładu

$h_{MS,\cdot}$ parametr wygładzania wyznaczony za pomocą zasady maksymalnego wygładzania

h_{LSCV} parametr wygładzania wyznaczony metodą kroswalidacyjną najmniejszych kwadratów

h_{PLCV} parametr wygładzania wyznaczony metodą kroswalidacyjną pseudowiarygodności

h_{BVC} parametr wygładzania wyznaczony metodą obciążoną kroswalidacyjną

h_{SCV} parametr wygładzania wyznaczony metodą wygładzoną kroswalidacyjną

h_{PM} parametr wygładzania wyznaczony metodą podstawiania

h_{UH} parametr wygładzania wyznaczony metodą uogólnionej średniej harmonicznej

$\hat{f}(\mathbf{x}, \mathbf{H})$ estymator jądrowy wielowymiarowej funkcji gęstości

$K^P(\mathbf{x})$ produktowa funkcja jądra

$K^R(\mathbf{x})$ radialna funkcja jądra

$\mathbf{H}$ macierz parametrów wygładzania

$|\mathbf{H}|$ wyznacznik macierzy parametrów wygładzania

Wprowadzenie

Uniwersalne a jednocześnie skuteczne procedury statystyczne stanowią jedno z podstawowych narzędzi w szeroko pojętych badaniach ekonomiczno-społecznych. Potrzeba stosowania takich procedur jest ściśle związana z charakterem zjawisk dotyczących wysoko rozwiniętego społeczeństwa, gdzie stopień różnorodności wszelkich aspektów jego funkcjonowania jest wysoki.

W analizach statystycznych związanych ze zjawiskami ekonomicznymi szczególną rolę odgrywają metody nieparametryczne. Zapewniają uniwersalność ze względu na brak konieczności przyjmowania dodatkowych założeń dotyczących rozkładów zmiennych losowych. Założenia takie nie zawsze bowiem są spełnione, co w wielu przypadkach jest wyrazem szczególnego charakteru analizowanych zmiennych ekonomicznych – ich wyjątkowości i niepowtarzalności. Zatem, niepowszedniość i incydentalność zmiennych ekonomicznych wymusza aplikację takich metod, dla których obserwowany jest szeroki zakres stosowalności, przy jednoczesnym braku konieczności przyjmowania, być może, wątpliwych założeń. Z drugiej strony dobre własności procedur nieparametrycznych stanowią gwarancję skuteczności metod nieparametrycznych.

Funkcja gęstości jest jedną z podstawowych charakterystyk zmiennej losowej, stąd nieparametryczne procedury związane z funkcją gęstości odgrywają coraz większą rolę w analizach zmiennych ekonomicznych. Nieparametryczna estymacja funkcji gęstości w wielu przypadkach jest nie tylko procedurą stanowiącą punkt wyjścia do dalszych szczegółowych analiz statystycznych dotyczących zmiennej losowej, ale stanowi również niezwykle zwartą i wyczerpującą procedurę dostarczającą określone spektrum informacji o własnościach zmiennej losowej. Różnorodność rodzajów tego podejścia oznacza, z jednej strony, wykorzystanie wyników estymacji funkcji gęstości do określenia w sposób jednoznaczny klasy procedur wykorzystywanych w dalszych badaniach i analizach statystycznych, z drugiej zaś może stanowić docelową i kompletną procedurę statystyczną. Jest zatem procedurą statystyczną o bardzo ogólnym charakterze, jednocześnie gwarantując szczegółowość wyników analiz.

Jądrowa estymacja funkcji gęstości jest procedurą stosowaną w analizach statystycznych nie tylko dotyczących zjawisk ekonomicznych, ale również przyrod-

niczych czy technicznych, jednak konieczność podjęcia decyzji o dwóch istotnych parametrach metody jądrowej, tj. funkcji jądra i parametru wygładzania, w dużym stopniu ogranicza stosowalność tej metody. Wielość propozycji metod dostępnych w literaturze, związanych z wyborem odpowiednich parametrów metody jądrowej, prowadzi do znacznego uproszczenia stosowalności metody jądrowej funkcji gęstości. W większości przypadków wykorzystywane są te metody wyboru parametrów metody jądrowej, które są łatwe do implementacji, co często oznacza niedoskonałość tych procedur.

Tematem niniejszej monografii są procedury wyboru parametrów metody jądrowej, ze szczególnym uwzględnieniem parametru wygładzania, gdyż w literaturze przedmiotu wskazuje się na mniejsze znaczenie wyboru funkcji jądra w estymacji jądrowej funkcji gęstości. Wybór tematyki jest naturalnym rozwinięciem i podsumowaniem wcześniejszych badań autorki w tym zakresie. Dotyczyły one, w szczególności, analiz związanych z wyborem funkcji jądra, wyborem parametru wygładzania, stosowalności procedur jądrowych w badaniach ekonomicznych, co wiązało się z odpowiednią modyfikacją klasycznych procedur jądrowej estymacji funkcji gęstości, uwzględniających charakter tych zmiennych.

Głównym celem badawczym autorki monografii jest analiza własności procedur wyboru parametru wygładzania w jądrowej estymacji funkcji gęstości, przy czym zasadniczym atrybutem branym pod uwagę przy analizach badań związanych z parametrem wygładzania był nie tylko aspekt metodologiczny, ale przede wszystkim możliwość zastosowania w badaniach zmiennych ekonomicznych. Analizie poddano zarówno klasyczne, jak i nieklasyczne procedury wyboru parametru wygładzania, wskazując w ten sposób na procedury optymalne w określonym zagadnieniu badawczym. Przedstawiono również wyniki badań dotyczących zaproponowanej autorskiej metody wyboru parametru wygładzania, uwzględniającej informacje dodatkowe związane z populacją.

Realizacja powyższego celu badawczego wymagała sprecyzowania celów szczegółowych, które są sformułowane następująco:

- określenie rodzajów funkcji jądra, ich szczegółowa prezentacja oraz wskazanie możliwości zastosowania określonej klasy funkcji jądra w konkretnej sytuacji badawczej, biorąc pod uwagę informacje dodatkowe dotyczące rozkładu zmiennej losowej,
- prezentacja metod wyboru wartości parametru wygładzania wraz z podaniem najważniejszych własności metod,
- porównanie wartości parametrów wygładzania w estymacji jądrowej funkcji gęstości, dokonane przy uwzględnieniu różnych postaci funkcji jądra, co prowadzi do wskazania najlepszych par parametrów metody jądrowej w estymacji funkcji gęstości,
- analiza własności autorskiej metody wyboru parametru wygładzania,
- określenie najlepszych metod wyboru parametru wygładzania dla estymacji funkcji gęstości zmiennej losowej w badaniach ekonomicznych.

W wyniku analiz związanych z badaniem własności metod wyboru parametru wygładzania w estymacji jądrowej funkcji gęstości weryfikacji podlegały następujące hipotezy badawcze:

- wybór parametru wygładzania w estymacji jądrowej funkcji gęstości zmiennej losowej w badaniach ekonomicznych jest ściśle uzależniony od informacji dodatkowych dotyczących podstawowych charakterystyk rozkładu, takich jak asymetryczność i wielomodalność,
- w określonych sytuacjach badawczych możliwe jest wskazanie optymalnych par parametrów metody jądrowej poprzez wskazanie najlepszej funkcji jądra dla określonej wartości parametru wygładzania,
- przy estymacji jądrowej funkcji gęstości zmiennej losowej w badaniach ekonomicznych informacje dodatkowe dotyczące jej rozkładu upraszczają procedurę wyboru optymalnych parametrów metody jądrowej z punktu widzenia określonego błędu,
- modyfikacje klasycznych procedur estymacji funkcji gęstości poprzez dobór nieklasycznych postaci funkcji jądra oraz zastosowanie nieklasycznych metod wyboru parametru wygładzania powodują polepszenie własności estymacji jądrowej zmiennej losowej w badaniach ekonomicznych.

Dla potrzeb realizacji celów dobrano odpowiednią strukturę pracy. W rozdziale pierwszym przedstawione są informacje związane z estymacją nieparametryczną, w szczególności z estymacją jądrową funkcji gęstości. Podane są własności estymatora klasycznego typu Rosenblatta-Parzena, jak również zaprezentowane są modyfikacje postaci klasycznej estymatora wynikające z wystąpienia ograniczonego nośnika zmiennej losowej, co w badaniach ekonomiczno-społecznych występuje dość często. W rozważaniach ogólnych dotyczących klasycznego podejścia w estymacji jądrowej szczególnego znaczenia nabiera szczegółowe określenie miary precyzji estymacji. W wielu metodach wyboru parametru wygładzania w estymacji jądrowej funkcji gęstości minimalizacja określonej ściśle miary precyzji pozwala na wyznaczenie właściwej wartości parametru wygładzania. Rozwinięciem rozważań związanych z klasycznym podejściem jest prezentacja procedur dotyczących estymacji pochodnych określonego rzędu funkcji gęstości.

W rozdziale drugim podjęto próbę usystematyzowania informacji dotyczących funkcji jądra prezentowanych w literaturze przedmiotu. Wyodrębniono osiem rodzajów funkcji jądra, gdzie czynnikami klasyfikującymi odpowiednią funkcję jądra do określonej grupy była konstrukcja, zasady stosowania w analizach praktycznych, charakter zmiennych losowych, których funkcja gęstości jest objęta estymacją oraz optymalność funkcji jądra. Wyodrębnione w ten sposób klasy w wielu przypadkach nie są rozłączne.

Rozdział trzeci jest poświęcony metodom wyboru parametru wygładzania. Różnorodność prezentowanych w literaturze przedmiotu metod powoduje pewną nieokreśloność co do metody najlepszej. Stąd, w rozdziale tym przedstawiono zarówno te metody, które są traktowane przez badaczy jako proste i szybkie

(uwypuklając ich wady), jak i metody bardziej zaawansowane, które mimo trudności w implementacji powinny być stosowane w praktyce. Zaprezentowano również autorską metodę wyboru parametru wygładzania opartą na uogólnionej średniej harmonicznej. Uzupełnieniem tych rozważań jest porównanie wybranych metod wyboru parametru wygładzania, ze szczególnym uwzględnieniem zależności między funkcją jądra i parametrem wygładzania oraz przy uwzględnieniu, przy wyborze metody, informacji dodatkowej związanej ze zmienną losową.

W rozdziale czwartym omówiono metody estymacji wielowymiarowej funkcji gęstości oraz procedury wyboru macierzy parametrów wygładzania. Odmienność metod wyboru parametrów w estymacji funkcji gęstości, przedstawionych w tym rozdziale, wynika bezpośrednio z faktu, że procedury stosowane dla wielowymiarowej estymacji nie są jedynie prostym rozszerzeniem procedur jednowymiarowych.

W rozdziale piątym wyniki analiz zawartych w poprzednich rozdziałach zostały wykorzystane w zagadnieniach związanych z estymacją funkcji gęstości zmiennej losowej w badaniach ekonomicznych, uwzględniając w sposób szczególny informacje wstępne dotyczące charakteru rozważanej zmiennej losowej.

Przy przygotowywaniu publikacji opierano się głównie na literaturze anglojęzycznej, gdyż w literaturze polskiej jedynie kilka opracowań dotyczy problematyki nieparametrycznej estymacji jądrowej funkcji gęstości. Monografia zatem może stanowić uzupełnienie tej luki.

W części pracy związanej z zastosowaniem procedur oraz z analizami porównawczymi stosowano metody symulacyjne przy wykorzystaniu oprogramowania MATLAB firmy Mathworks, wersja R2012a i R2014a.

Autorka pragnie serdecznie podziękować Panu Profesorowi zw. dr. hab. Czesławowi Domańskiemu za życzliwość i wsparcie oraz Recenzentowi – Panu Profesorowi dr. hab. Grzegorzowi Kończakowi za cenne uwagi i sugestie zawarte w recenzji, które wpłynęły z pewnością na poprawę jakości publikacji.

Rozdział 1

Estymacja nieparametryczna funkcji gęstości

1.1. Uwagi wstępne

Termin „statystyka parametryczna" bezpośrednio i jednoznacznie wskazuje na procedury związane z charakterystyką (parametrem) populacji, które stosowane są na podstawie dostępnych danych, na przykład eksperymentalnych. Parametr może być rozumiany w dwojaki sposób: jako niesprecyzowana stała występująca w rodzinie rozkładów zmiennej losowej lub też, wykorzystując określenie w szerszym sensie, parametr może oznaczać prawie wszystkie metody opisu zmiennej losowej w określonej rodzinie rozkładu (Gibbons, Chakraborti, 2003). Uwzględniając powyższe określenie statystyki parametrycznej, statystyka nieparametryczna rozumiana może być jako zbiór procedur, albo ściśle nieparametrycznego rodzaju (na przykład test nieparametryczny oznaczający weryfikację hipotezy, która nie dotyczy wartości parametru), lub też procedury stanowiące analogię do klasycznego (parametrycznego, uwzględniającego arbitralne założenie postaci badanych funkcji) podejścia, gdzie określone założenia dotyczące rozkładu są zastąpione przez założenia o bardziej ogólnym charakterze niż w przypadku podejścia klasycznego. Mimo że procedury parametryczne charakteryzowane są prostotą teoretyczną i obliczeniową oraz powszechną znajomością i dostępnością w literaturze, nie są one wystarczalne w wielu sytuacjach badawczych.

Procedury nieparametryczne są procedurami uniwersalnymi (Domański, 1979, 1990), w związku z czym mogą one być stosowane odnośnie do różnorodnych zagadnień poświęconych analizom populacji. Są wykorzystywane w celu identyfikacji rozkładu populacji, jak również służą do opracowania wniosków związanych ze szczegółową charakterystyką zmiennej losowej w populacji. Ich uniwersalność ma również odzwierciedlenie w możliwości stosowania procedur nieparametrycznych bez konieczności przyjmowania konkretnych założeń o populacjach, z których otrzymujemy dane rzeczywiste, co jest sytuacją wymuszoną w praktyce, gdy brak jest informacji wstępnej o rozkładzie badanej populacji lub istnieje duże ryzyko związane z przyjęciem założenia dotyczącego rozkładu

(Domański i in., 1998). W wielu przypadkach są one z jednej strony łatwiejsze do implementacji, a z drugiej charakteryzują się jedynie nieznacznie mniejszą skutecznością niż procedury parametryczne. Należy jednak podkreślić, że w wielu przypadkach procedury nieparametryczne, chociaż koncepcyjnie proste oraz przejrzyste w zakresie interpretacji, wymuszają konieczność zastosowania odpowiedniej techniki komputerowej, co stanowiło wyraźną barierę zarówno w badaniach teoretycznych, jak i aplikacyjnych.

Określenie „nieparametryczne metody statystyczne" zostało wprowadzone do terminologii statystycznej przez Wolfowitza w 1942 roku (Wolfowitz, 1942) i związane było z koniecznością rozszerzenia stosowalności metod statystycznych ponad zwyczajowo wówczas wykorzystywane metody parametryczne.

Należy zauważyć, że propozycje teoretyczne dotyczące metod nieparametrycznych oraz próby aplikacji pojawiały się już na początku XVIII wieku. John Arbuthnot, w pracy wykorzystującej ewidencję chrztów dzieci w Londynie w latach 1629–1710, z podziałem na płeć (Arbuthnot, 1710), analizował źródło zaobserwowanych regularności i upatrywał je w opatrzności boskiej (Ostasiewicz, 2012). Jednocześnie była to pierwsza próba zastosowania testu znaków, co przez niektórych statystyków (Noether, 1984; Domański, 1986) traktowane jest również jako prezentacja pierwszego testu statystycznego weryfikującego hipotezy statystyczne. Idea „wszechobecnego bóstwa", zapewniającego określone wartości średniej statystycznej, jest utożsamiana z podwalinami rozwoju osiemnastowiecznej statystyki.

Jednak to prace z początku XX wieku są traktowane jako właściwe początki dziedziny znanej jako statystyka nieparametryczna, w szczególności prace Pearsona (1900, 1911) dotyczące zgodności rozkładów, praca Hotellinga i Pabsty z 1936 roku dotycząca korelacji rang (Hotelling, Pabst, 1936) oraz praca Wilcoxona (1945) poświęcona testom Wilcoxona dla jednej i dwóch prób.

Akceptacja terminologii wprowadzonej przez Wolfowitza dotycząca statystyki nieparametrycznej nie była powszechna. Działo się to pomimo tego, że Wolfowitz, obok konieczności prowadzenia badań w tym nowym wówczas obszarze statystyki, prezentował również próbę zastosowania zasady ilorazu wiarygodności w przypadku nieparametrycznym, którą Neyman i Pearson zaproponowali już 10 lat wcześniej dla przypadku parametrycznego. W latach 40. XX wieku jedynie nieliczni statystycy z Uniwersytetu Columbia oraz Uniwersytetu Princeton wykorzystywali to określenie w publikacjach w „Annals of Mathematical Statistics". Natomiast w „Journal of the American Statistical Associations" po raz pierwszy termin nieparametryczny pojawił się dopiero w 1949 roku.

Praca Scheffègo (1943) jest próbą nie tylko przedstawienia teoretycznych podstaw do rozwoju statystyki nieparametrycznej, ale przede wszystkim jest pierwszą publikacją prezentującą w kompletny sposób istniejące dotychczas nieparametryczne metody statystyczne, w tym nieparametryczne metody weryfikacji hipotez statystycznych (testy zgodności, testy losowości, testy dla dwóch prób, testy niezależności i testy analizy wariancji). Praca ta jest traktowana jako pionierska

w zakresie zdefiniowana estymacji nieparametrycznej, uwypuklając potencjalne problemy związane z terminologią. Scheffè, w celu uniknięcia niejednoznaczności związanej z nazewnictwem, proponuje θ (będące parametrem rozkładu) nie nazywać parametrem, lecz jedynie liczbą rzeczywistą określoną przez rozkład. W zagadnieniach związanych z estymacją punktową wskazana została jedynie równoważność estymatorów parametrycznych i nieparametrycznych w zakresie nieobciążoności i zgodności. Problemy estymacji przedziałowej przedstawione w równie wąskim zakresie związane były jedynie z przedziałami ufności dla mediany, dla różnicy dwóch median oraz przedziałami ufności dla nieznanej dystrybuanty.

Pierwsze próby estymacji charakterystyk funkcyjnych zostały sformalizowane i przedstawione w pracy Walda i Wolfowitza (1939), a dotyczyły one obszaru ufności dla dystrybuanty.

Rozwój metod nieparametrycznych w ostatnich czterdziestu latach jest ściśle związany z rozwojem technik obliczeniowych. Zwiększenie mocy obliczeniowej współczesnych komputerów umożliwiło szybki rozwój metod nieparametrycznych, w tym zwiększenie liczby propozycji modyfikacji, mających na celu poprawę efektywności rozważanych metod.

Procedury nieparametryczne dotyczące charakterystyk funkcyjnych zmiennych losowych, na przykład estymacja funkcji gęstości, uwzględniają fakt, że zbiór funkcji określonych na zbiorze liczb rzeczywistych R jest nieporównywanie liczniejszy od R niż w przypadku na przykład estymacji liczby rzeczywistej lub wektora (Gajek, Kałuszka, 1996). Może zatem wystąpić sytuacja, że nawet po zawężeniu klasy możliwych funkcji gęstości do zbioru funkcji i-krotnie różniczkowalnych ($i \in$ N, N oznacza zbiór liczb naturalnych), nie istnieje estymator nieobciążony funkcji.

Podejście nieparametryczne w estymacji charakterystyk funkcyjnych zmiennej losowej umożliwia przyjęcie zdecydowanie słabszych założeń dotyczących postaci funkcyjnej charakterystyki podlegającej estymacji w porównaniu z przyjmowanymi założeniami w metodach parametrycznych. W podejściu parametrycznym wymagane jest przyjęcie założenia, że znana jest rodzina funkcji gęstości, z której pochodzą obserwacje, na przykład w estymacji funkcji gęstości model parametryczny zakłada, że funkcja gęstości jest znana co do skończonej liczby parametrów. Istnienie informacji wstępnej (określane jako „boskie spostrzeżenie") o postaci funkcyjnej charakterystyki podlegającej estymacji powinno być wykorzystane i wówczas zastosowanie procedur parametrycznych traktowane jest jako wskazane. Gdy natomiast takich informacji wstępnych brak i założenie związane z badaną charakterystyką funkcyjną oparte jest jedynie na niedostatecznych przesłankach lub brak jest takich przesłanek, procedura estymacji parametrycznej może prowadzić do nieprawidłowych wyników dotyczących charakterystyk funkcyjnych podlegających wnioskowaniu. Opierając estymację na założeniu, że funkcja gęstości jest określonym elementem ze znanej rodziny parametrycznej, badacz

musi brać pod uwagę możliwość błędnej specyfikacji modelu, co może oznaczać, że model nie jest zgodny z populacją, z której dane zostały pobrane. W szczególności gdy przyjmowane jest założenie o normalności, w rzeczywistości narzucona jest grupa całkiem restrykcyjnych założeń, na przykład dotyczących symetryczności, jednomodalności, określonej monotoniczności poza wartością modalną. Jeśli rzeczywista funkcja gęstości jest asymetryczna lub posiada wiele wartości modalnych, wówczas założenie o normalności może prowadzić do niewłaściwej charakterystyki funkcji gęstości i może skutkować fałszywymi wynikami estymacji oraz błędnym wnioskowaniem. Oczywiście możliwe jest postępowanie polegające na testowaniu, czy zakładany rozkład jest zgodny z rzeczywistością. Ale odrzucenie założenia dotyczącego rozkładu, niestety, nie skutkuje określeniem rozkładu alternatywnego, na przykład odrzucenie hipotezy o normalności sprawia, że badacz nie otrzymuje informacji o innym możliwym rozkładzie, zmuszony jest wrócić do punktu wyjścia.

Podejście nieparametryczne oznacza uniknięcie problemów związanych z koniecznością specyfikacji parametrycznej postaci funkcyjnej przed dokonaniem procedur estymacyjnych. Przyjmowane są jedynie założenia, że spełnione są pewne warunki regularności, takie jak gładkość i różniczkowalność. A to są zdecydowanie słabsze założenia dotyczące struktury postaci funkcyjnej funkcji gęstości niż w przypadku metod parametrycznych. Na przykład, w estymacji rozkładu dochodów zamiast przyjmowania założenia, że funkcja gęstości należy do rodziny rozkładów normalnych lub lognormalnych, przyjmuje się jedynie założenie, że funkcja gęstości jest dwukrotnie lub trzykrotnie różniczkowalna. Przyjęcie słabszych założeń dotyczących struktury postaci funkcyjnej gęstości, niestety, powoduje konieczność posiadania większej liczby danych by otrzymać ten sam stopień dokładności, co właściwie wyspecyfikowany model parametryczny. Podejście nieparametryczne może oznaczać zatem konieczność zwiększenia liczby obserwacji w procedurach nieparametrycznych, by osiągnąć taki sam poziom precyzji, jak odnośnie do dobrze wyspecyfikowanego modelu parametrycznego. Z drugiej jednak strony podejście nieparametryczne pozwala na większą elastyczność w stosowaniu procedur statystycznych, ponieważ wymagane jest jedynie założenie, że nieznana funkcja gęstości należy do pewnego nieskończonego zbioru krzywych.

Nieparametryczna estymacja funkcji gęstości w wielu przypadkach traktowana jest jako wstępny etap w analizie lub też jako analiza dotycząca dokładnie określonej charakterystyki funkcyjnej zmiennej losowej.

Hansen (2009) określa podejście parametryczne jako podejście o skończonej wymiarowości, podczas gdy podejście nieparametryczne to podejście o nieskończonej wymiarowości.

Zasadnicza różnica związana z procedurami estymacji dotyczy szybkości zbieżności. Podczas gdy w podejściu parametrycznym, właściwie wyspecyfikowanym, dla liczebności próby n, szybkość zbieżności jest rzędu $n^{-1/2}$, to w przypadku podejścia nieparametrycznego tempo to jest wolniejsze niż $n^{-1/2}$. Przy czym, w odróż-

nieniu od parametrycznego podejścia, tempo zbieżności jest zazwyczaj odwrotnie proporcjonalne do liczby zmiennych, co jest znane w literaturze jako „przekleństwo wymiarowości". Jest ono traktowane jako jedno z najważniejszych utrudnień pojawiających się w procedurach nieparametrycznych (Stone, 1994; Pagan, Ullah, 1999). Stosowanie prawie każdej procedury statystycznej jest powiązane, mniej lub bardziej, z przekleństwem wymiarowości, w podejściu nieparametrycznym oznacza ono jednak konieczność stosowania bardzo dużych prób w celu zapewnienia dokładności na odpowiednim poziomie. Metody nieparametryczne są zatem zalecane szczególnie wtedy, gdy liczba zmiennych jest mała, natomiast zbiór danych duży (Silverman, 1986). Stosownie metod nieparametrycznych oznacza konieczność rozważenia odpowiednio dużej liczebności próby w porównaniu z liczbą zmiennych.

Ponadto w podejściu parametrycznym nie istnieje rozróżnienie między prawdziwym modelem występującym w rzeczywistości a modelem wykorzystywanym w procedurze estymacji, natomiast w podejściu nieparametrycznym taka różnica między modelami istnieje.

Metody nieparametryczne powodują wyższy stopień skomplikowania modelu dopasowywanego w zależności od próby. Im więcej informacji w próbie (co może oznaczać większą liczebność próby), tym większy stopień złożoności modelu. Wymaga to odrębnych twierdzeń dotyczących rozkładów.

W podejściu nieparametrycznym modele dopasowywane są traktowane jako aproksymacje i dlatego też są z góry skazane na błędną specyfikację, a to implikuje obciążenie estymatora. Zazwyczaj wzrost złożoności dopasowywanego modelu powoduje zmniejszenie obciążenia, ale oznacza jednocześnie wzrost wariancji estymacji. Stosowanie metod nieparametrycznych oznacza zatem konieczność uwzględniania tego kompromisu, co powoduje ustalenie takiego stopnia złożoności modelu, by zminimalizować ogólne miary dopasowania, na przykład błąd średniokwadratowy (*MSE*).

Nieparametryczne procedury mogą dotyczyć, między innymi, funkcji gęstości (jednowymiarowej i wielowymiarowej), pochodnych funkcji gęstości, warunkowych funkcji gęstości, dystrybuanty, funkcji regresji, parametrów położenia, w tym mediany i kwantyli, parametrów skali, w tym wariancji. Nieparametryczna estymacja charakterystyk funkcyjnych może być głównym celem podjętych badań, ale może również stanowić punkt wyjściowy będący warunkiem, którego spełnienie umożliwia dalsze drugoetapowe zagadnienia estymacji lub weryfikacji hipotez statystycznych. Jeżeli problem rozważany w drugim etapie dotyczy parametrów (skończonego wymiaru), wówczas estymacja traktowana jest jako semiparametryczna. Badacz nie specyfikuje postaci funkcyjnej dla pewnych zakresów danych, ale dla niektórych zakresów założenia parametryczne są konieczne.

Nieparametryczne metody obejmują pewien rodzaj aproksymacji oraz metody wygładzania (jądrowe, szeregów, splajnów). Nieparametryczne metody są indeksowane poprzez parametr wygładzania (parametr dostrajania), który określa

stopień złożoności. Wybór tego parametru wygładzania jest najczęściej zagadnieniem kluczowym w zastosowaniach metod w praktyce. Oznacza to, że metody określania parametru wygładzania oparte na danych są traktowane jako istotne w nieparametrycznych metodach. Natomiast konieczność ich określenia jest często uważana za wadę metod nieparametrycznych. Metody nieparametryczne, które wymagają parametru wygładzania, ale nie mają określonej zależnej od danych reguły wyboru parametru wygładzania, są traktowane jako niekompletne (Hansen, 2009). Niestety, takie podejście występuje dość często, co jest związane z trudnościami w rozwijaniu szczegółowych zasad i reguł wyboru parametru wygładzania. Jest to pewien kompromis, gdyż parametr wygładzania jest wybierany w oparciu o odpowiedni problem statystyczny.

Funkcja gęstości jest jedną z podstawowych charakterystyk funkcyjnych zmiennej losowej. Warto pokreślić zarówno jej znaczenie teoretyczne, jak i wykorzystanie do praktycznych obliczeń i ilustracji wyników wieloaspektowych analiz (Kulczycki, 2005). Funkcja gęstości jest stosowana między innymi do wyznaczenia prawdopodobieństwa, że zmienna losowa przyjmuje wartość z ustalonego zbioru, natomiast jej prezentacja graficzna stanowi wygodne i intuicyjne narzędzie wstępnej analizy danych. W modelowaniu statystycznym funkcja gęstości stanowi opis schematu losowej zmienności danych, które nie są wyjaśnione przez inne strukturalne charakterystyki w modelu (Bowman, Azzalini, 2004). W praktycznych zagadnieniach funkcja gęstości najczęściej nie jest znana i musi być oszacowana na podstawie danych pochodzących z próby.

Estymacja funkcji gęstości umożliwia wyjaśnienie i ocenę, czy potencjalny model jest dopasowany do danych rzeczywistych. Pagan i Ullah (1999) zwracają uwagę na znaczenie estymatora funkcji gęstości w analizach Monte Carlo dotyczących estymatorów podlegających analizie. Nieparametryczny estymator funkcji gęstości dostarcza całościowy obraz rozkładu estymatora i dlatego jest stosowany jako wygodna forma prezentacji wyników eksperymentów Monte Carlo. Przykładowo, estymator funkcji gęstości jest konieczny, gdy parametryczny estymator ma rozkład asymptotyczny, który zależy od gęstości szacowanej w określonym punkcie.

Estymacja funkcji gęstości może być wykorzystywana jako zasadniczy i podstawowy etap w analizie danych, o ile celem analizy statystycznej jest otrzymanie, na przykład, dogodnej formy prezentacji struktury danych. Może to być związane z analizą wartości modalnych, co prowadzi do wskazania odmiennych aspektów rozważanego zagadnienia, a następnie umożliwia dekompozycję przedmiotową. Analiza ogonów estymatora funkcji gęstości, rozważanych, na przykład, w procesach produkcyjnych, gdzie „lekki" ogon wskazuje na zwykły rozrzut technologiczny, „ciężki" ogon – na zużycie sprzętu i konieczność wymiany konkretnych podzespołów (Kulczycki, 2005), również ma znaczenie praktyczne. W przypadku zmiennej losowej wielowymiarowej estymator jądrowy wielowymiarowej funkcji gęstości umożliwia analizę struktury danych w zakresie zależności między poszczególnymi współrzędnymi tej zmiennej.