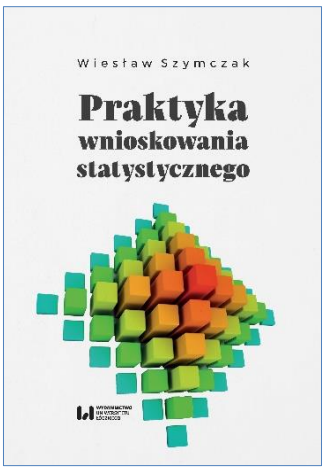
Wiadomości Statystyczne. The Polish Statistician, 2019, vol. 64, 8, 61–68
Informacje. Recenzje. Dyskusje / Information. Reviews. Discussions

Wiesław Szymczak
Praktyka wnioskowania statystycznego
(Wydawnictwo Uniwersytetu Łódzkiego, Łódź 2018)



Statystyka współcześnie służy właściwie wszystkim dziedzinom nauki. Jej intensywny rozwój oraz specyfika poszczególnych gałęzi badawczych pociągnęły za sobą wypracowanie zróżnicowanych preferencji w zakresie stosowania metod i narzędzi statystycznych oraz ich doskonalenia w określonych kierunkach. Przekłada się to – rzecz jasna – m.in. na oryginalność w zakresie pragmatyki wnioskowania statystycznego. Niemniej jednak interpretacja danych oraz wyników analiz, a także ocena przydatności poszczególnych rozwiązań w konkretnych sytuacjach i formułowanie konkluzji opierają się na uniwersalnych regułach warsztatu statystycznego i wymaga odpowiednich umiejętności oraz doświadczenia.

Omawiana książka jest kolejną, po ciekawej publikacji pióra Ostasiewicza (2012), pracą poświęconą właściwościom myślenia statystycznego – w tym ograniczeniom, jakie można napotkać, stosując narzędzia statystyczne – oraz problemom i błędom występującym we wnioskowaniu statystycznym. Odpowiednio do zakresu zainteresowań badawczych i kompetencji autora opracowanie koncentruje się na specyfice stosowania statystyki w naukach psychologicznych i naukach o zdrowiu. W przedmowie autor słusznie podkreśla, że aby można było posłużyć się określoną metodą statystyczną i udowodnić dane twierdzenia, konieczne jest spełnienie odpowiednich założeń, o co korzystający z metod statystycznych nie zawsze należycie dbają. Oprócz tego zasadnie zauważa występowanie *kultu istotności statystycznej* – opieranie się głównie na wyznaczonej ocenie istotności zmiennej czy funkcji parametrów, co niejako zwalnia badacza z myślenia. Dlatego też używanie zaawansowanego oprogramowania przez osoby bez doświadczenia w statystyce prowadzi nierzadko do bezsensownych wniosków lub przyzwolenia na to, by procedury podejmowały decyzje za badacza.

Książka składa się ze wstępu oraz z pięciu rozdziałów opisujących podstawowe paradygmaty i poszczególne elementy wnioskowania statystycznego. Autor rozpoczyna od wprowadzenia rozróżnienia między statystyką jako dyscypliną naukową – przez którą rozumie statystykę matematyczną (nazywając ją też statystyką teoretyczną lub teorią statystyki) – a zastosowaniem metod statystycznych w praktyce. Można mieć co do tego rozróżnienia pewne wątpliwości. Po pierwsze, statystyka jako nauka obejmuje także kierunki i sposoby wykorzy-

stania oferowanych przez siebie narzędzi oraz analizę ich użyteczności. Po drugie, autor uważa, że statystyka teoretyczna ma charakter dedukcyjny, natomiast praktyczna – indukcyjny. Nie do końca można się z tym zgodzić, ponieważ w wymiarze praktycznym dedukcja też ma szerokie zastosowanie, np. dedukcyjne metody imputacji i edycji danych są powszechnie znane i bardzo często stosowane w badaniach statystycznych (zob. np. de Waal, Pannekoek i Scholtus, 2011). Nieco nieprecyzyjnie ujęto też kwestię powstania geometrii nieeuklidesowej. U jej podwalin legły bezowocne próby wyprowadzenia V postulatu Euklidesa na podstawie czterech pozostałych. Nikołaj Łobaczewski i János Bolyai nie zmienili V postulatu, podali tylko jego zaprzeczenie, a następnie na bazie tego zaprzeczenia i pozostałych postulatów stworzyli rzeczoną geometrię. Jednak zaproponowana przez nich opcja nie jest jedynym możliwym zaprzeczeniem V postulatu – w XVIII w. alternatywne rozwiązania podali m.in. Giovanni Girolamo Sacchieri i Johann Heinrich Lambert (zob. np. Kordos, 1994). Natomiast interesująco autor wprowadza czytelnika w tematykę swojej książki, zaznaczając rozbieżności między twierdzeniami teoretycznymi i ich wartością a możliwościami ich zastosowań w praktyce.

Rozdział 1 zawiera ciekawą i wszechstronną dyskusję dotyczącą istoty wnioskowania statystycznego, paradygmatów statystycznych oraz najważniejszych teorii testowania hipotez. Autor poświęca sporo uwagi rozbieżnościom między poszczególnymi definicjami prawdopodobieństwa oraz podejściami Fishera i Neymana-Pearsona do konstrukcji testów statystycznych. Bardzo słusznie dostrzega przy tym różnicę między metaanalizą a analizą danych z konkretnego badania oraz częste ignorowanie konsekwencji zastosowania estymacji w wynikach analiz. Nasuwają się tu jednakże pewne kwestie dyskusyjne.

Po pierwsze, autor w dość zawiły sposób przedstawia interpretację poziomu istotności *ex post* (zwanego także *wartością p* lub – z języka angielskiego – *p-value*) testu statystycznego. Najczęściej nazywa je bowiem „prawdopodobieństwem w teście". Nie bardzo wiadomo, jak należałoby to rozumieć. Sformułowanie takie zdaje się sugerować, jakoby w teście było zagnieżdżone jakieś bliżej nieokreślone prawdopodobieństwo. Przytoczono też kilka objaśnień tego terminu z literatury przedmiotu dotyczącej zastosowań statystyki w psychologii i medycynie, np. „zdarzyło się coś nielosowego" (przy $p < 0{,}05$), „zaszło zdarzenie prawie niemożliwe" (gdy $p < 10^{-6}$) czy też „wartość *p* pozwala określić, jak przekonująco dane świadczą przeciwko hipotezie zerowej o losowości" (trudno zresztą powiedzieć, jak formalnie wygląda owa „hipoteza o losowości"). Sam autor zauważa skądinąd, że tego rodzaju interpretacje są nieprzejrzyste, jednak brakuje wyraźnego wskazania, że poziom istotności *ex post* oznacza najmniejszy poziom istotności, na którym następuje odrzucenie hipotezy zerowej przy danej wartości testu. Zatem np. jeśli $p = 0{,}03$, to wiadomo, że na poziomie istotności *ex ante* (tj. zakładanym arbitralnie) równym 0,05 brak jest podstaw do odrzucenia hipotezy zerowej, natomiast jeżeli poziom ten wynosi 0,02, to hipote-

zę zerową się odrzuca. Jest to czytelne objaśnienie, zrozumiałe także dla osób niezbyt zaawansowanych statystycznie. Natomiast całkowicie słusznie autor zauważa, że wiązanie poziomu istotności *ex post* z zależnością (choć nie jest jasne, między czym a czym) nie ma żadnego uzasadnienia statystycznego – nawet jeśli wziąć pod uwagę modele regresyjne, o których zresztą w tym kontekście nie wspomina. Nie każdy test jest bowiem powiązany z zależnością między zmiennymi czy wielkościami.

Po drugie, autor rozumie wnioskowanie statystyczne jako postępowanie wykorzystujące metody statystyczne umożliwiające uogólnienie zależności zaobserwowanych w próbie na populację generalną, z której ta próba pochodzi. Takie postrzeganie tego zagadnienia wydaje się nazbyt zawężone, ponieważ wnioskowanie statystyczne może przebiegać wielotorowo i dotyczyć np. konkluzji płynących z analizy skupień, analizy regresji czy analizy czynnikowej.

Po trzecie, autor w nieco nadmiernie uproszczony sposób traktuje lemat Neymana-Pearsona, prezentując go w formie *de facto* parametrycznej. Tymczasem jego najbardziej ogólna wersja ma postać nieparametryczną, ponieważ odnosi się do identyczności rozkładu danej zmiennej z rozkładem P_0 w hipotezie zerowej oraz z rozkładem P_1 w hipotezie alternatywnej – zob. np. Zieliński (1990) czy Bartoszewicz (1996). Zabrakło też podkreślenia, że test wskazany w lemacie jest testem jednostajnie najmocniejszym ze wszystkich testów na danym poziomie istotności *ex ante*. Dziwi to tym bardziej, że sam autor nieco wcześniej celnie zauważa, że w praktyce mało się dba o moc używanego testu. O testach jednostajnie najmocniejszych mowa jest dopiero później, podczas gdy należałoby napomknąć o tym już w trakcie prezentacji teorii Neymana-Pearsona. W odniesieniu do parametrycznych hipotez złożonych autor raz podaje, że poziom istotności *ex ante* zależy od statystyki testowej (φ), a następnie, że tylko od parametrów określonych weryfikowanymi hipotezami (θ).

Autor słusznie zauważa, podzielając pogląd innych badaczy, że żaden pracownik naukowy nie może jednego ustalonego poziomu istotności używać w każdym czasie i we wszystkich okolicznościach, by odrzucać hipotezę. Nie można jednak zapomnieć, że poziom istotności *ex post* właśnie w tym pomaga. Warto też nadmienić niebłahą rolę przyzwyczajenia do określonych poziomów istotności *ex ante*, np. 0,05. Wątpliwości można mieć z kolei co do sugerowanej przez autora niekompatybilności teorii Fishera i Neymana-Pearsona. W obu przypadkach mamy do czynienia bezpośrednio z testowaniem hipotezy zerowej, w obu przypadkach występuje też *de facto* poziom istotności *ex post* – choć w teorii Neymana-Pearsona pośrednio, jako graniczny próg poziomu istotności *ex ante* dla odrzucenia hipotezy zerowej. Tym bardziej zatem nie ma powodu, aby teorię Fishera nazywać testowaniem istotności (bo właściwie istotność czego się tutaj testuje?).

Na zakończenie rozdziału 1 autor podaje przykład ilustrujący zagrożenie wynikające z nadmiernego opierania się na istotności korelacyjnej kosztem meryto-

rycznej oceny potencjalnych zależności. Pewną pomocą w tym zakresie mogłoby być wyznaczenie mocy testu, o co jednak w programach statystycznych raczej trudno – być może dlatego, że zależy ona od konkretnej hipotezy alternatywnej.

W rozdziale 2 autor wskazuje różnice pomiędzy rozmaitymi programami komputerowymi w zakresie rezultatów zastosowania zaimplementowanych narzędzi statystycznych (w tym testów) oraz możliwości weryfikacji założeń tych narzędzi. Czyni to na przykładzie empirycznych danych dotyczących efektywności trzech terapii leczenia „łokcia tenisisty", używając analizy wariancji. Już na początku przypomina – co być może stanowi truizm, o którym jednak warto nie zapominać – że hipoteza zerowa w testach jednorodności winna być prawdziwa. Co więcej, założenie to jest konieczne w praktycznie każdym teście do ustalenia teoretycznego rozkładu statystyki testowej. Z kolei w przypadku normalności rozkładu autor uwypukla możliwy dysonans pomiędzy normalnością rozkładu w próbce a normalnością rozkładu w populacji, sygnalizując, że próbka może wskazywać na nienormalny rozkład populacji, podczas gdy faktycznie jest on normalny i na odwrót. Warto byłoby wspomnieć tutaj o centralnym twierdzeniu granicznym. Jeśli jego założenia (niezależność i jednakowy rozkład obserwacji) są spełnione, to rozkład sum obserwacji dostatecznie licznej próbki bardzo dobrze przybliża normalny rozkład populacji, nawet gdy ta próbka jest nominalnie losowana z rozkładu całkiem innego niż normalny.

Zaprezentowane w tym i w następnych rozdziałach obliczenia autor przeprowadza za pomocą trzech programów statystycznych: SPSS 24, STATA 13 i SYSTAT 13, ponieważ, jak twierdzi, dysponuje licencją na te programy. Trochę szkoda, że nie uwzględnił w tym kontekście także niektórych przynajmniej środowisk typu *open source*, a zatem całkowicie bezpłatnych, jak choćby środowisko R (R Core Team, 2019) – największe i najwszechstronniejsze, ciągle rozwijane ogólnoświatowe oprogramowanie do analiz statystycznych składające się z wielu specjalistycznych pakietów. W celu szerszej oceny efektywności programów w tym zakresie można by też rozważyć przeprowadzenie stosownej symulacji np. metodą bootstrapową z danymi wyjściowymi wylosowanymi generatorem liczb losowych z odpowiednio dobranego rozkładu teoretycznego.

Jak już wspomniano, do pełnej wiedzy i rzetelnego podjęcia decyzji na podstawie testu statystycznego niezbędna staje się znajomość jego mocy. Rozdział 3 pokazuje główne problemy, jakie się z tym wiążą. Przede wszystkim dotyczy to hipotezy alternatywnej. W przypadku hipotez złożonych za moc testu przyjmuje się maksymalne prawdopodobieństwo odrzucenia hipotezy zerowej, gdy jest ona fałszywa. Problem polega na tym, że to maksimum może nie istnieć, a wtedy trzeba szacować moc testu asymptotycznie. Poza tym, wedle autora, nie ma czegoś takiego jak empiryczna moc testu. Niektóre programy komputerowe obliczają wprawdzie moc testu na podstawie danych empirycznych, ale trudno powiedzieć, przy użyciu jakiego algorytmu. Jednak np. w środowisku R znane są

co najmniej dwa kierunki wyznaczania mocy testu: analityczny (funkcje power.anova.test, power.prop.test i power.t.test pakietu stats oraz pakiet pwr – zob. Blomberg, 2014) i symulacyjny, oparty na funkcji elastycznej (pakiet paramtest – zob. Hughes, 2017 i Wang, 2014). Autor analizuje rozpatrywane przez siebie programy pod względem różnic między nimi odnośnie do szacowania mocy testu i wielkości próby. Opiera się przy tym na bogatych danych empirycznych. Twierdzi jednak, że hipoteza alternatywna jest dopełnieniem hipotezy zerowej do całego zbioru hipotez możliwych do sformułowania w analizowanym zagadnieniu. Nie do końca jest to prawda, ponieważ w testach występują także alternatywne hipotezy jednostronne, niekoniecznie wyczerpujące zbiór potencjalnych opcji w tym zakresie (np. dla $H_0: \theta = \theta_0$ może być $H_1: \theta < \theta_0$).

Rozdział 4 autor poświęca podobnej analizie, tyle że dotyczącej oceny wielkości efektu, która ma doprecyzować wynik testu i określić jego jakość. Ocena ta ma swe źródło w teorii Neymana-Pearsona, a jej mierników używa się niemal wyłącznie w psychologii oraz naukach medycznych i o zdrowiu. Autor mocno i zasadnie krytykuje spotykany w literaturze pogląd, że interpretacja wielkości efektu jest niezależna od tego, jakie zmienne (i jak mierzone) są rozpatrywane oraz o jaki efekt chodzi. Na podstawie danych empirycznych wskazuje na podobieństwa i odmienności w zakresie realizacji i wielkości współczynników d Cohena i η^2 oraz sygnalizuje sprzeczności występujące między wynikiem testowania a wnioskowaniem na podstawie mierników wielkości efektu, podkreślając, że to drugie nie powinno zastępować pierwszego. Można tu sfomułować następujące spostrzeżenia:

- niezbyt jasne wydaje się stwierdzenie autora, że w teorii Neymana-Pearsona hipoteza zerowa jest hipotezą prostą w postaci „którą wolimy odrzucić"; dlaczego właściwie równość dwóch parametrów lub hipotetyczną wartość parametru mielibyśmy z góry deprecjonować?;
- autor utrzymuje, że interpretacja miernika wielkości efektu jako oceny siły zależności oznacza najpierw porównywanie statystyki będącej podstawą testu (z czym?), a następnie porównywanie prawdopodobieństw odpowiadających tym wartościom – o jakie prawdopodobieństwa tutaj chodzi: popełnienia błędu II rodzaju czy o *p*-value?;
- bardzo trafna jest uwaga, że w przypadku gdy program komputerowy poda prawdopodobieństwo wielkości 0,000, to nieprawdą jest, że wynosi ono 0 – okazuje się jedynie mniejsze od 0,0005; podobnie zresztą w publikacjach statystycznych 0 oznacza, że odnośna wartość była mniejsza niż 0,5, a 0,0 – że wyniosła mniej niż 0,05 – jest to jeden z fundamentów właściwego posługiwania się statystyką;
- nie bardzo wiadomo, jaki dokładnie mechanizm regresji krokowej zastosował autor w swej analizie; ukazany przebieg algorytmu nie wskazuje ani na podejście *forward*, ani na *backward* – może była to opcja *stepwise* (*bidirectional*)?;

- klasyfikacja wielkości efektu podana w postaci pojedynczych liczb (prawdopodobnie wartości progowych) jest myląca – w takiej sytuacji należy podawać konkretne i dokładne przedziały wartości, tak jak uczyniono to zresztą w odpowiednich tablicach;
- nie wyjaśniono, czym są wielkości $\mu_1, \mu_2, \ldots, \mu_k$ w jednoczynnikowej analizie kowariancji, skoro μ oznacza wyraz wolny, stały dla modelu;
- przydałoby się wyjaśnić, jaką postać kontrastów ortogonalnych zastosowano w przypadku testu Mauchly'ego; jego statystyka tego właśnie wymaga;
- autor słusznie zauważa, że współczynniki determinacji R^2 szacowane w modelach regresji logistycznej dla różnych zmiennych wynikowych z badania nie mogą być bezpośrednio porównywane, a porównywanie R^2 w modelu regresji logistycznej i liniowej (klasyczna metoda najmniejszych kwadratów, OLS) jest problematyczne; warto byłoby zatem wspomnieć w tym miejscu, że w regresji liniowej w ogóle nie można porównywać R^2 dla modeli z różnymi zmiennymi objaśniającymi (np. Y i $\ln Y$), nawet jeżeli zmienne objaśniające są identyczne;
- pojęcia *zmienne wynikowe z badania* używało się dużo wcześniej, niż zostało zdefiniowane jego znaczenie; zdaniem autora w przypadku regresji logistycznej określenie to jest lepsze niż *zmienna zależna*;
- autor dostrzega, że z rachunków nie bardzo wiadomo, jakie znaczenie z medycznego punktu widzenia mają oszacowane ryzyka wystąpienia chorób układu krążenia u badanych osób; i słusznie, bo interpretacja takich informacji należy już do finalnych użytkowników posiadających ekspercką wiedzę z tej dziedziny, a nie do statystyków;
- nieprawdziwe jest stwierdzenie, że „możemy w tym momencie uznać, że decyzje w wersji parametrycznej i nieparametrycznej są równoważne"; równość rozkładów (którą autor rozważa) implikuje oczywiście równość ich parametrów (np. wartości oczekiwanych czy wariancji), ale nie na odwrót: identyczność (np. wartości oczekiwanych) nie pociąga za sobą tożsamości rozkładów;
- nie jest jasne, po czym liczona jest pochodna statystyki będącej podstawą testu Manna-Whitneya, obliczana w programie SPSS, i jak ta pochodna wygląda.

W ostatnim rozdziale autor zajmuje się innymi podejściami do wnioskowania statystycznego opartymi na paradygmatach: bayesowskim i wiarygodnościowym. Korzystając z przykładowych danych, wskazuje zalety opcji bayesowskiej, przede wszystkim uzyskiwanie całego spektrum rozkładu ryzyka zamiast pojedynczego oszacowania jego wielkości. Interesujący jest tutaj zwłaszcza model kancerogenezy. Szkoda jedynie, że mankamenty prezentacji zmniejszają czytelność opisu. Na przykład autor podaje dwie, jego zdaniem, alternatywne postaci funkcji dawka-odpowiedź, które jednak są… identyczne. Zabrakło również praktycznej interpretacji tej funkcji. Warto też zauważyć, że określenie *percentyl* w naszej nomenklaturze odnosi się zazwyczaj do kwantyli rzędu $n/1000$, $n = 1, 2, \ldots, 999$,

a kwantyle rzędu $k/100$, $k = 1, 2, \ldots, 99$ określa się mianem centyli[1]. Drugą część omawianego rozdziału stanowią rozważania na temat paradygmatu wiarygodnościowego i opartego na nim testowania. Obejmują one m.in. konstrukcje oraz właściwości funkcji wiarygodności. Czasami jednak widać tutaj problematyczne kwestie. Przykładowo autor raz określa funkcję wiarygodności wektora parametrów w formie klasycznej $L(\theta; x)$, by za chwilę uczynić to w formie warunkowej $L(\hat{\theta}|x)$ – a to przecież nie to samo. Ponadto nie wiadomo, co oznacza „prawdopodobieństwo odpowiedzi" $P^*(d; \theta)$. Dwie, wedle autora, alternatywne postaci opartej na funkcji wiarygodności funkcji $\lambda(x)$, czyli λ i λ^*, są takie same. Poza tym, jeśli rozważamy najlepsze wyjaśnienie w hipotezie alternatywnej jako odpowiedni kres górny funkcji wiarygodności, to ten kres winien być wyznaczany po obszarze odrzuceń hipotezy zerowej, a nie po całym zbiorze możliwych wartości badanego parametru. W przykładzie dotyczącym testowania hipotez autor twierdzi, że wartość funkcji wiarygodności dla domniemanego parametru powinna znajdować się w liczniku wyrażenia określającego iloraz wiarygodności, by za chwilę podać je tam w mianowniku.

Podsumowaniem książki jest krótki przegląd zaczerpniętych z literatury odpowiedzi na pytania na temat działań statystycznych, w rodzaju „Co powinienem zrobić?", „Co «mówią» dane?", „W co powinienem wierzyć?". Wydaje się, że zakończenie powinno być nieco szersze i zawierać przede wszystkim syntetyczny zestaw wniosków płynących z przeprowadzonych rozważań. Natomiast za bardzo przydatny należy uznać zamieszczony na końcu książki słowniczek najważniejszych stosowanych w niej pojęć.

W całej książce występują drobne usterki redakcyjno-techniczne, które zmniejszają nieco jej przejrzystość. Reasumując, należy jednak pokreślić, że *Praktyka wnioskowania statystycznego* stanowi bardzo wartościowy przyczynek do dyskusji na temat właściwego stosowania metod statystycznych. Dzięki temu może spełnić ona ważną funkcję w edukacji statystycznej. Należy także żywić nadzieję, że autor będzie podążał w obranym kierunku i rozciągnie swoje wnikliwe rozważania na kolejne istotne oraz doskonalone teoretyczne i informatyczne narzędzia statystyki.

Andrzej Młodak
(Państwowa Wyższa Szkoła Zawodowa im. Prezydenta Stanisława Wojciechowskiego w Kaliszu)

[1] Natomiast w nomenklaturze anglojęzycznej kwantyle rzędu $k/100$, $k = 1, 2, \ldots, 99$ to *percentiles* (od angielskiego *percent* – procent), skąd zapewne wynika ta rozbieżność.

BIBLIOGRAFIA

Bartoszewicz, J. (1996). *Wykłady ze statystyki matematycznej*. Warszawa: Wydawnictwo Naukowe PWN.

Blomberg, S. P. (2014). *Power Analysis Using R*. Semantic Scholar. Pobrane z: https://pdfs.semanticscholar.org/2b85/0bb035e93663f835a5453c4f02ae55ac65da.pdf.

Hughes, J. (2017). *Simulating Power with the paramtest Package*. The Comprehensive R Archive Network (R–CRAN). Pobrane z: https://cran.r-project.org/web/packages/paramtest/vignettes/ Simulating-Power.html.

Kordos, M. (1994). *Wykłady z historii matematyki*. Warszawa: Wydawnictwa Szkolne i Pedagogiczne.

Ostasiewicz, W. (2012). *Myślenie statystyczne*. Warszawa: Oficyna Wydawnicza Wolters Kluwer business Polska.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Pobrane z: https://www.R-project.org/.

de Waal, T., Pannekoek, J., Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*. New Jersey: John Wiley & Sons. DOI: 10.1002/9780470904848.

Wang, C. (2014). Asymptotic power of likelihood ratio tests for high dimensional data. *Statistics & Probability Letters*, *88*, 184–189.

Zieliński, R. (1990). *Siedem wykładów wprowadzających do statystyki matematycznej*. Warszawa: Państwowe Wydawnictwo Naukowe.