Lublin, dnia 23.12.2020

dr hab. Oleg Gorbaniuk

Instytut Psychologii KUL

Katolicki Uniwersytet Lubelski Jana Pawła II

Recenzja rozprawy doktorskiej Lilianny Jarmakowskiej-Kostrzanowskiej pt. Klasyczne oraz bayesowskie testowanie hipotez na przykładzie porównań dwóch grup w badaniach psychologicznych

Rozprawa doktorska napisana pod kierunkiem prof. dr hab. Tytusa Sosnowskiego

Promotor pomocniczy: dr Rafał Wieczorek

Celem recenzowanej rozprawy było porównanie podejścia klasycznego (testowanie istotności statystycznej hipotezy zerowej) z podejściem bayesowskim na bazie danych generowanych komputerowo z rozkładu teoretycznego (zawężonego do rozkładu ex-gaussowskiego) oraz rozkładu empirycznego (zawężonego do rozkładu umiarkowanie skośnego) pochodzącego ze wcześniejszych badań. Przedmiotem oceny była zgodność i niezgodność decyzji podejmowanych w procesie wnioskowania statystycznego oraz poziom popełnianych błędów I i II rodzaju w obu podejściach w schemacie badawczym przewidującym porównywanie dwóch grup/prób niezależnych. W badaniach manipulowano wielkością efektu (w przypadku rozkładu teoretycznego) oraz wielkością próby (w przypadku rozkładu teoretycznego i empirycznego). W przypadku rozkładu teoretycznego efekty manipulacji sprawdzano dodatkowo w trzech warunkach wiedzy apriori badacza: intuicji badacza zgodnej i niezgodnej z rzeczywistością oraz w warunkach niewiedzy. Specyfika postawionych problemów badawczych nie wymagała zbierania danych empirycznych

specjalnie w celu ich rozstrzygnięcia, stąd uwzględniono dane ze źródeł wtórnych – już istniejących – które były podstawą wcześniejszej publikacji (Jarmakowski-Kostrzanowski, Jarmakowska-Kostrzanowska, 2016).

Zasadniczo podstawowa część tytułu rozprawy precyzyjnie odzwierciedla treść pracy, natomiast podtytuł nawiązujący do badań psychologicznych ma bardziej umowny charakter. Rozprawa doktorska dotyczy metodologii analizy danych i nie jest specyficzna wyłącznie dla psychologii: wnioski z badań mogą być generalizowane na wszystkie dziedziny nauk, gdzie mamy do czynienia z wnioskowaniem statystycznym na bazie prób pobieranych z populacji o rozkładzie skośnym.

Ocena struktury pracy

Praca składa się z trzech części: wprowadzenia teoretycznego, omówienia badań własnych i dyskusji, przy czym w ramach pierwszych dwóch części pojawia się w sumie dziewięć rozdziałów o wspólnej numeracji, z kolei dyskusja jest z niej wyłączona. Praca zawiera ponadto streszczenie, wykaz cytowanej literatury i załączniki. Praca ma 168 numerowanych stron, przy czym numeracja zaczyna się po spisie treści.

Mam kilka zastrzeżeń co do sposobu ustrukturalizowania treści pracy biorąc pod uwagę standardy pisania prac w psychologii. Przede wszystkim oczekuje się wyraźnego rozdzielenia opisu metody badań od analizy danych/wyników. W aktualnej wersji pracy funkcje „metody" pełnią (pod)rozdziały 6.2 (nieujęty w spisie treści), 7, początek 8 („Opis procedury symulacji") i 9.1-9.2, z kolei funkcję „wyników" pełnią podrozdziały 8.1.-8.2, gdzie zawarta jest analiza danych pochodzących z rozkładu teoretycznego oraz podrozdział 9.3, który zawiera analizę danych z rozkładu empirycznego. Przy czym rozdział 8 i podrozdział 9.2 mają taki sam tytuł („Opis procedury symulacji"), ale zupełnie inny zasięg treści. Istnieje przynajmniej kilka możliwych rozwiązań przy takiej serii badań, aby struktura

raportu z badań nabrała większej przejrzystości metodologicznej. W pracy jest też sporo nienumerowanych podrozdziałów (np. ss. 64, 72, 76), a tekst jest przeciążony przypisami dolnymi.

Czas od czasu w pracy pojawiają się drobne błędy we wzorach (np. s.60), błędy językowe, pominięcia słów lub przeczenia diametralnie zmieniające znaczenie zdania (zob. streszczenie), ale przy ponownym przeczytaniu zdania można się domyślić jego właściwego znaczenia. Nie ma też numeracji wzorów wg APA.

Ocena wprowadzenia teoretycznego pracy

We wprowadzeniu teoretycznym Autorka bardzo dobrze opisała podejście klasyczne (rozdział 1) i podejście bayesowskie (rozdział 2) w testowaniu hipotez statystycznych. Sposób omówienia, użyte przykłady i metafory wskazują na dobre rozumienie przez Autorkę istoty obu podejść. Komunikatywność i przystępność treści dla przeciętnego czytelnika kwalifikowałyby tę część pracy do publikacji jako rozdziału podręcznika z zakresu metodologii nauk społecznych. Jest to wiedza ważna z punktu widzenia tematu rozprawy, ale ma charakter wprowadzający i porządkujący. Rozdziałem, który bezpośrednio koresponduje z tematem pracy, jest rozdział 3 („Klasyczne i bayesowskie testowanie hipotez statystycznych"), do którego tytułu należałoby dodać „porównanie", aby odróżnić od poprzednich dwóch rozdziałów (1 i 2) o podobnej nazwie. W tym rozdziale czytelnik oczekiwałby ukazania aktualnego stanu wiedzy opartej na wynikach analiz porównawczych obu podejść na bazie rozkładów teoretycznych i/lub empirycznych. Obszerniejszy przegląd badań – mniej lub bardziej odległych od schematu porównania dwóch grup/prób – pozwoliłby zrealizować kilka celów: ukazać lukę w dotychczasowej wiedzy, uzasadnić obiektywną nowość i potrzebę nowych badań, omówić typowe schematy postępowania i metody stosowane w tego typu badaniach etc. Aktualnie rozdział przywołuje trzy badania (dwa z

1987 i jedno z 2017 roku), w tym tylko jedno – autorstwa Jeon i De Boecka (2017) – zostało szczegółowo omówione. Jest ono kluczowe, ponieważ stanowi bardzo bliski odpowiednik badań zrealizowanych później przez Autorkę, w tym również w zakresie zastosowanej metody, co decyduje o ich kompatybilności. Zasadnicza różnica dotyczy zakresu generalizacji: Jeon i De Boeck (2017) ograniczyli swoje analizy do zmian wartości p i czynnika Bayesa BF_{10} w przypadku rozkładu normalnego w populacji. To ograniczenie stanowiło dla Autorki główny argument na rzecz rozszerzenia badań na inne typy rozkładów, odbiegające od rozkładu normalnego. Warto odnotować, że literatura cytowana przez Autorkę w rozprawie kończy się na 2018 roku. Z uwagi na kluczowe znaczenie rozdziału 3 warto byłoby poszerzyć go o omówienie innych badań porównujących podejścia w innym schemacie niż testowanie różnic pomiędzy dwoma grupa/próbami. Pozwoliłoby to wychwycić wspólne/powtarzające się prawidłowości w zachowaniu wartości p i czynnika Bayesa BF_{10}, stanowiące następnie podstawę – obok teorii – dla sformułowania hipotez badawczych. Dostarczyłoby to też dodatkowych punktów odniesienia w ocenie wyników badań własnych podczas ich dyskusji.

Ocena empirycznej części pracy (badań własnych Autorki)

W swojej rozprawie Autorka nie formułuje hipotez badawczych, sprowadzając cele badań do eksploracyjnych (z tego punktu widzenia tytuł rozdziału 5 „Cel badań i hipotezy badawcze" jest nieuzasadniony), a pytania mają charakter opisowy. Wydaje się, że przynajmniej część przypuszczeń możliwa była do sformułowania w postaci hipotez badawczych, których uzasadnienia można byłoby doszukiwać się zarówno w teorii (np. dysponowanie w podejściu bayesowskim odpowiedzią „niewiadomo/niekonkluzywność danych" w przedziale $1/3 < BF_{10} < 3$) jak i w bliźniaczych (ale heterozygotycznych) badaniach Jeon i De Boecka (2017). Dokonany przeze mnie pobieżny przegląd dostępnej

literatury wskazuje, że podjęty przez Autorkę problem badawczy jest obiektywnie nowy, a przedstawione przez Autorkę uzasadnienie wskazuje także, że ten problem jest również ważny z punktu widzenia metodologii badań w psychologii.

Autorka szczegółowo wyjaśniła wszystkie elementy i etapy zastosowanej metody badań bez większych skrótów myślowych. Na marginesie: można byłoby rozważyć wybór poziomu mocy testu $1 - \beta = 0{,}90$, który jest aktualnie częściej stosowany jako progowy, ale nie miałoby to większego znaczenia dla uzyskanych wyników i wniosków z badań.

Wyniki analiz danych generowanych z rozkładu teoretycznego lub pobieranych z rozkładu empirycznego w celu przetestowania wpływu wielkości próby i/lub wielkości efektu na poziom błędu I i II rodzaju oraz wskaźniki zgodności/niezgodności podejmowanych decyzji w ramach porównywanych podejść nie budzą zastrzeżeń. Analiza została wykonana zgodnie z przyjętymi założeniami, a interpretacja wyników jest adekwatna do zawartości tabel. Analizę danych przedstawioną w tabelach w ramach głównego tekstu rozprawy uzupełniają wyniki analiz przedstawione w załącznikach dla innych (pośrednich) poziomów wielkości efektu (0,1, 0,3, 0,4, 0,6, 0,7) niż te, które wybrane jako progowe (0,2, 0,5, 0,8). Przeprowadzone przez Autorkę analizy wykazały, że stopień zgodności decyzji podejmowanych w podejściu klasycznym i bayesowskim zależy od wielkości efektu, wielkości próby i rodzaju prawdopodobieństwa apriorycznego, czego należałoby spodziewać się biorąc pod uwagę czynniki wpływające na moc testu w podejściu klasycznym oraz konstrukcję czynnika Bayesa BF_{10}. Jak już wcześniej wspomniałem, zabrakło tu bardziej zdecydowanej postawy Autorki w procesie formułowania odpowiedzi na wszystkie stawiane pytania badawcze (1a-1c). Wyniki analiz pozwoliły ponadto ustalić, przy jakiej kombinacji wielkości efektu i wielkości próby osiąga się pełną zgodność pomiędzy konkurującymi ze sobą podejściami oraz w jakim przedziale mniejszych prób w ramach danej wielkości efektu podejście bayesowskie ujawnia swoją wyższość nad podejściem klasycznym z uwagi m.in. na

oferowanie badaczowi możliwości oceny stopnia popierania hipotezy przez zebrane dane (pod warunkiem użycia przez badacza informatywnego prawdopodobieństwa apriorycznego reprezentującego uprzednią wiedzę badacza). W przypadku nieinformatywnego prawdopodobieństwa apriorycznego przy małych próbach podejście klasyczne z kolei wypada lepiej z uwagi na niższy błąd II rodzaju.

Każda z dwóch podstawowych części analiz, odpowiadających danym generowanym z rozkładu teoretycznego i danym z rozkładu empirycznego, jest zakończona dyskusją (podrozdziały 8.1.1.3, 8.1.2.2.4 i 8.2.2.4). Z tego powodu dyskusja ogólna na końcu rozprawy znacznym stopniu dubluje treści już zaprezentowane/przedyskutowane. Nowym elementem jest w niej natomiast omówienie przyszłych kierunków badań (w sumie 5-6), których inspiracją są wyniki badań własnych. Jednym z celów każdej dyskusji jest konfrontacja wyników badań własnych z dotychczasowym stanem wiedzy wynikającym z literatury przedmiotu, a w szczególności oczekuje się porównania wyników z wynikami badań bliźniaczych. Z uwagi na bardzo wąski krąg badań uwzględnionych w ramach takiego przeglądu w rozdziale 3 oczekiwałbym tu szczegółowej konfrontacji z wynikami badań Jeon i De Boecka (2017), czego Autorka nie zrobiła. Stanowi to okazję do postawienia pytań: W jakim stopniu wyniki badań przedstawione w pracy – oceniając z perspektywy czasu – były do przewidzenia na podstawie badań Jeon i De Boeck (2017) i teorii? Czy przyjęty kryterium $BF_{10} > 3$ odpowiada pod względem stopnia konserwatyzmu poziomowi błędu I rodzaju 0,05 (czy są one względnie równoważne w badaniach własnych Autorki)? Przy jak dużej skośności rozkładu należy oczekiwać znacznej rozbieżności pomiędzy wynikami porównań podejść dla rozkładu normalnego i nie-normalnego? Jakie rozbieżności w zachowaniu wartości p i czynnika Bayesa BF_{10} są do przewidzenia w przypadku porównywania podejścia klasycznego i bayesowskiego w innych schematach badań niż testowanie różnicy dwóch prób

niezależnych i czy istnieje potrzeba takich porównań (w szczególności uwzględniając nowsze wyniki analogicznych badań innych autorów)?

Wnioski końcowe

Mimo zgłoszonych uwag uważam, że Autorka recenzowanej rozprawy jest przygotowana do samodzielnego prowadzenia pracy naukowej, wykazuje się obszerną wiedzą metodologiczną, a jej rozprawa wnosi wartościowy wkład w metodologię nauk społecznych, w tym psychologii. W szczególności przyczynia do zmiany sposobu myślenia o statystyce i poszerza naszą wiedzę na temat możliwości i ograniczeń podejścia bayesowskiego względem klasycznego. Tym samym stwierdzam, że rozprawa doktorska Lilianny Jarmakowskiej-Kostrzanowskiej spełnia wymogi stawiane pracy doktorskiej (Ustawa z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym) i wnioskuję o jej dopuszczenie do dalszych etapów przewodu doktorskiego.

dr hab. Oleg Gorbaniuk