

XXVII Konferencja Naukowa „Zastosowania statystyki i data mining w badaniach naukowych”

The 27th Scientific Conference ‘Applications of statistics and data mining in scientific research’

19 października 2023 r. odbyła się (w formule online) XXVII Konferencja Naukowa „Zastosowania statystyki i data mining w badaniach naukowych” organizowana przez firmę StatSoft Polska. Patronat nad nią objęło Polskie Towarzystwo Statystyczne, a patronami medialnymi zostały czasopisma: „Wiedza i Życie”, „Świat Nauki”, „Forum Akademickie” i „Wiadomości Statystyczne. The Polish Statistician”.

Uczestnicy tego popularyzującego wiedzę statystyczną wydarzenia mogli się dowiedzieć, jak skutecznie wydobywać informacje ze zbiorów danych i efektywnie prezentować uzyskane wyniki. Po raz kolejny konferencja stała się platformą wymiany myśli i doświadczeń oraz wydarzeniem integrującym środowisko zajmujące się analizą danych. Interdyscyplinarny charakter spotkania pozwolił uczestnikom szerzej spojrzeć na poprawne planowanie badania oraz dobór odpowiednich metod analizy danych empirycznych.

Wykładowcy, wśród których znaleźli się przedstawiciele polskich ośrodków naukowych i analitycy ze StatSoft Polska, omawiali sposoby wykorzystywania metod analizy danych w takich obszarach, jak: dobór prób w badaniach trudno dostępnych lub nieznanych zbiorowości ludzkich, zastosowanie nowoczesnych metod analizy skupień – DBSCAN i OPTICS (rozwijające tradycyjne metody aglomeracyjne), analiza statystyczna wyników poniżej granicy wykrywalności („danych niewykrytych”), praktyczna i statystyczna istotność wyników oraz ich znaczenie w poprawnej ocenie rezultatów badań, odszyfrowywanie złożonych biomarkerów w badaniach wielkoskalowych, a także zwiększenie skuteczności oddziaływania na odbiorców dzięki prezentowaniu wyników analiz statystycznych w przystępny sposób.

W pierwszym wystąpieniu *Czy jeśli czegoś nie znajdziemy, to czy na pewno tego nie ma? Sposoby analizy statystycznej danych niewykrytych* dr hab. Łukasz Binkowski, prof. UKEN (Uniwersytet Komisji Edukacji Narodowej w Krakowie) wskazywał, jak radzić sobie w analizie statystycznej z przypadkami wyników poniżej granicy wykrywalności, występującymi w badaniach empirycznych (zwłaszcza w naukach ścisłych). Badacze mają z nimi do czynienia w sytuacjach, gdy przy użyciu zaawansowanej aparatury mierzą pewne parametry z wielką dokładnością. I mimo że w ramach rutynowych analiz mierzą na przykład stężenie różnych substancji na poziomie jednej części na miliard, to i tak – z uwagi na bardzo niskie stężenia w materiałach badawczych – bardzo często nie udaje się wykryć niektórych substancji. Oficjal-

nie takie przypadki określa się mianem *wyników poniżej granicy wykrywalności* lub – w zależności od przyjętego kryterium – *wyników poniżej granicy oznaczalności*. Prelegent odniósł się do często pojawiających się wątpliwości związanych z tym tematem: Co taki wynik właściwie oznacza? Czy mierzonego analitu po prostu w próbce nie ma, czy może jest, ale w tak małej ilości, że aparatura nie jest w stanie go wykryć? Co należy zrobić, gdy w projekcie pojawiają się takie przypadki? Jak je traktować w analizie statystycznej i czy wypracowano w odniesieniu do nich specjalne procedury statystyczne? Omówił również koncepcję granic raportowania i błędne praktyki, a przede wszystkim wskazał sposoby poprawnego ujęcia tych granic w analizach statystycznych.

W referacie *Coś więcej niż metody aglomeracyjne – współczesne metody analizy skupień w praktyce* mgr Anna Wilk (kierownik Działu Analiz StatSoft Polska) przedstawiła nowoczesne metody analizy: DBSCAN (ang. *density-based spatial clustering of applications with noise*) i OPTICS (ang. *ordering points to identify the clustering structure*), wskazała różnice między nimi i omówiła ich zastosowania. Analiza skupień (nazywana również segmentacją) to jedna z najpopularniejszych metod analizy danych. Polega na wyszukiwaniu i wyodrębnianiu z danych skupień, czyli grup podobnych obiektów. Zastosowanie analizy skupień jest bardzo szerokie – od projektów naukowych i badawczych po segmentację klientów, grupowanie dokumentów i analizę obrazów. Takie nowoczesne techniki, jak DBSCAN i OPTICS oferują znacznie większą elastyczność i skuteczność w wyodrębnianiu z danych skupień o różnych kształtach (np. podłużnym czy wklęsłym) niż tradycyjne metody aglomeracyjne. Prelegentka dodatkowo wyjaśniła, na co należy zwracać szczególną uwagę przy przeprowadzaniu analizy z użyciem omawianych technik.

Dr hab. Sławomir Pasikowski, prof. UŁ (Uniwersytet Łódzki) wygłosił referat *Dobór śniegowy i inne łańcuchowe doборы próby w badaniach społecznych*. Dobór śniegowy należy do powszechnie stosowanych sposobów pobierania prób w badaniach społecznych. Związany jest z rozwojem zaawansowanych matematycznie podejść łańcuchowych, stanowiących odpowiedź na problem trudno dostępnych lub nieznanymi zbiorowości ludzkich. Prelegent zauważył, że pozornie nieskomplikowana procedura stanowi źródło popularności tej metody i powstawania jej karykaturalnych postaci. Przybliżył ponadto oryginalną podstawę teoretyczną doborów łańcuchowych, główne konsekwencje jej ewolucji, a także zagadnienie reprezentatywności próby i wyznaczania jej wielkości.

Pomiędzy wykładami uczestnicy mogli uzyskać od prelegentów odpowiedzi na zadawane w czasie wystąpień pytania i brali udział w quizach statystycznych z nagrodami. Zostali także zaproszeni do udziału w organizowanym corocznie przez firmę StatSoft Polska konkursie na najlepszą pracę doktorską i magisterską przygotowaną z zastosowaniem programu Statistica. Z okazji jubileuszowej XXV edycji

konkursu sekretarz komisji konkursowej dr Janusz Wątroba (dyrektor ds. dydaktyki i współpracy naukowej StatSoft Polska) zaprosił do rozmowy prof. dr. hab. Andrzeja Sokołowskiego z Uniwersytetu Ekonomicznego w Krakowie, przewodniczącego komisji konkursowej, dr Małgorzatę Misztal z Uniwersytetu Łódzkiego, dwukrotną laureatkę konkursu – w kategorii prac magisterskich i prac doktorskich, oraz prof. dr. hab. n. med. Wojciecha Fendlera z Uniwersytetu Medycznego w Łodzi, promotora prac doktorskich nagrodzonych w konkursie. Goście opowiadali o doświadczeniach związanych z pracą w komisji i wrażeniach z udziału w rywalizacji. Zachęcali, aby spróbować swoich sił w konkursie, w którym czekają atrakcyjne nagrody (dla autora, promotora i wydziału uczelni).

Następnie Janusz Wątroba w referacie *Jakie znaczenie ma praktyczna i statystyczna istotność wyników badania?* zwrócił uwagę, jak ważnym zagadnieniem dla poprawnej oceny rezultatów badań jest rozróżnienie pomiędzy istotnością statystyczną a praktyczną otrzymanych wyników. Jest to kluczowe w przypadkach, kiedy badania są przeprowadzane na bardzo dużych albo skrajnie małych zbiorach danych. Skutecznym sposobem uniknięcia problemów związanych z prawidłową interpretacją wyników jest podjęcie odpowiednich działań już na etapie planowania badań. Okazuje się jednak, że w określonych sytuacjach można również próbować zastosować podejście typu *a posteriori*. Wykładowca zilustrował prezentowane rozwiązania przykładami praktycznymi.

Wystąpienie dr. hab. inż. Jacka Pietraszka, prof. PK (Politechnika Krakowska) *Planowanie doświadczeń – ku prostocie metod i prezentacji wyników* było poświęcone potrzebie przedstawiania analiz i wyników w przystępny sposób, co powinno się przełożyć na zwiększenie skuteczności oddziaływania na odbiorców. Statystyka stosowana (w tym planowanie doświadczeń) bardzo często korzysta z narzędzi matematycznych i rachunku prawdopodobieństwa. Wspomniane metody bywają dość skomplikowane zarówno w warstwie teoretycznej, jak i obliczeniowej, a więc ich zastosowanie wymaga wykorzystania takich programów, jak Statistica. Utrudnia to jednak pracę przeciętnemu użytkownikowi, który stosuje określone metody, nie do końca znając ich założenia i jedynie powierzchownie rozumiejąc mechanizm ich działania, a interpretację wyników przeprowadza na podstawie utartych schematów i rozmaitych – jak określił je prelegent – „magicznych liczb” w rodzaju 0,05. Dla odbiorców analiz, którzy nie są biegli w matematyce i statystyce, takie surowe wyniki i bazujące na nich wnioskowanie mogą być trudne do zrozumienia. Prowadzi to często do sceptycyzmu lub ignorowania wyników dostarczonych analiz. Autor referatu podkreślił, że celem badaczy powinno być zwiększenie skuteczności oddziaływania na odbiorców dzięki prezentacji analiz i wyników w przystępny sposób. Opracowując analizy, raporty i wytyczne, warto pamiętać o słynnym stwierdzeniu Rogera Penrose’a, że każdy wzór matematyczny umieszczony w książce zmniejsza

o połowę liczbę jej czytelników. Referent przedstawił kilka prostych metod obliczeniowo-graficznych i technik prezentowania wyników, które znacznie ułatwiają odbiór przygotowywanych analiz.

Na zakończenie dr Konrad Stawiski (Uniwersytet Medyczny w Łodzi) w wystąpieniu *Odszyfrowywanie złożonych biomarkerów w badaniach wielkoskalowych – doświadczenia* podzielił się swoimi doświadczeniami dotyczącymi przeprowadzania badań biomarkerowych. Nowoczesna medycyna rozwija się w kierunku leczenia spersonalizowanego. Dzięki stosowaniu technik omicznych w badaniach biomarkerowych, np. sekwencjonowaniu nowej generacji, możliwe jest otrzymanie wielu informacji na temat pacjentów i próbek. Ze względu na wysokie koszty stosowania takich oznaczeń badacze są jednak często zmuszeni pracować na zbiorach danych z liczbą przypadków dużo mniejszą niż liczba zmiennych, a klasyczną analizę utrudniają m.in. konieczność wielokrotnego testowania hipotez czy efekt serii. Prelegent zaprezentował oprogramowanie OmicSelector, stworzone na podstawie wieloletnich doświadczeń Zakładu Biostatystyki i Medycyny Translacyjnej Uniwersytetu Medycznego w Łodzi w przeprowadzaniu badań biomarkerowych oraz tworzeniu testów prognostycznych i predykcyjnych, do którego dostęp jest otwarty. Omówił techniki modelowania, zjawiska przeuczenia, selekcji zmiennych i optymalizacji hiperparametrów, a także zastosowanie głębokich sieci neuronowych do ekstrakcji zmiennych głębokich oraz nauczanie przenoszone.

Prezentacje wygłoszone podczas konferencji i nagrania wystąpień można znaleźć na stronie internetowej organizatora www.statsoft.pl/konferencja.

Agnieszka Jabłońska

StatSoft Polska