

WIADOMOŚCI STATYSTYCZNE

CZASOPISMO GŁÓWNEGO URZĘDU STATYSTYCZNEGO
I POLSKIEGO TOWARZYSTWA STATYSTYCZNEGO

MIĘDZYNARODOWY ROK STATYSTYKI 2013
KONFERENCJA NAUKOWA
STATYSTYKA — WIEDZA — ROZWÓJ

Czesław DOMAŃSKI

Wyzwania wobec statystyki jako nauki

*„Aby pojąć Boga musimy
studiować statystykę, ponieważ są to
wskaźniki Jego zamiarów”*

*Florence Nightingale
(1820—1910)*

Statystyka została wydzielona jako odrębna dyscyplina. Jest metodą wydobywania informacji z zaobserwowanych danych. Służy podejmowaniu logicznych decyzji w warunkach niepewności. Jako taka, wiedza statystyczna jest cenna dla ludzi wszystkich zawodów. Powszechne rozumienie statystyki jest znacznie ważniejsze niż rozumienie jakiegokolwiek innej dziedziny nauki (Domański, 2011).

Matematyka jest najstarszą dyscypliną przez nas uprawianą. Statystyka, mimo swojego niezbyt długiego panowania, poczynawszy od ukazania się książki Johna Graunta *Naturalne i polityczne obserwacje poczynione na biuletynach śmiertelności* (czyli od 1662 r.), ma bogaty rodowód sięgający odległej starożytności. Rodowód statystyki sięga XXXII w. p.n.e., gdyż plemię Ashipu, mieszkające między Eufratem a Tygrysem, zajmowało się głównie udzielaniem porad w zakresie ryzyka i niepewności oraz trudnych decyzji.

Wiele złożonych problemów naszego życia wyglądałoby prościej, gdyby przed podjęciem działań najpierw stawiać pytania i uzyskiwać właściwe informacje. Pytania uważane są często za kłopotliwe, gdyż wymagają analizy, myślenia i formułowania wniosków. Działania takie zabierają nam czas i energię. Mogą też prowadzić do niepożądanego dezorientacji i zdenerwowania. Wielu ludzi, aby tego uniknąć opiera się na mądrości innych albo rzuca się na „głębokie wody” w nieznane sytuacje. Opieranie się bez reszty na „mądrości” innych lub emocjonalne posunięcia mogą prowadzić do nieporozumień, złego wyboru momentu działania i pomyłek. Porady mogą być pomocne, ale raczej jako produkt odniesienia niż samowystarczalne podejście. Działania „na hura” to czysta spekulacja i hazard.

Te przykłady są dalece obiecujące i można powiedzieć, że nasze metody mają się dobrze. Ta sytuacja wydaje się pozornie paradoksalna. Nigdy wcześniej nie było większej potrzeby myślenia statystycznego. Jesteśmy otoczeni wyzwaniem różnorodnych banków danych (choć rzadko zgodnych z oczekiwaniami), które wymagają coraz lepszych metod statystycznych, algorytmów i modeli systemów przetwarzania.

WYZWANIA

Jak można określić oczekiwania wobec statystyki, które nazwaliśmy wyzwaniami?

Rozwój statystyki jako nauki będzie coraz szybszy z następujących powodów:

- 1) jest to nauka dla wszystkich,
- 2) rozumienie statystyki jest znacznie ważniejsze niż rozumienie jakiegokolwiek innej dziedziny nauki,
- 3) większość badań statystycznych oparta jest na próbie reprezentatywnej,
- 4) statystyka jest kluczową technologią do zdobywania i przetwarzania informacji koniecznych do podejmowania decyzji,
- 5) statystyka pełni rolę służebną wobec innych nauk i wobec człowieka.

Można się spodziewać, że metodologia statystyki będzie obejmowała:

- 1) wielkie zbiory danych — Big Data — wymagające zaawansowanej statystyki, zwłaszcza w procesie szybkich analiz danych,
- 2) inteligentne systemy biznesu,
- 3) technologie informatyczne,
- 4) wizualizację i analizy funkcjonalności danych ekonomicznych.

Każda konstrukcja ogólna, teoretyczna, jak np. twierdzenie matematyczne, choć piękne samo w sobie, nabiera jednak pełni rzeczywistości, gdy jest zastosowane w praktyce. Każde takie twierdzenie trwa w uśpieniu jak królewna z baśni czekając we śnie na wyzwoliciciela, który ją pobudzi do prawdziwego, pełnego istnienia, budząc jednocześnie zamek i całe królestwo. Wyzwolenie królewny powoduje statystyka.

jest zasadniczym wyzwaniem dla wszystkich statystyków. Możemy patrzeć na naszych wybranych liderów, instytucje, towarzystwa i oczekiwać od nich wkładni, ale poprawa na oczekiwaną skalę nastąpi tylko wtedy, jeśli wszyscy uczynimy z tego nasz priorytet.

Patrząc w przyszłość, odbudowa *image'u* musi być naszym głównym priorytetem. Musimy zrozumieć, że statystyka jest nauką niezbędną dla właściwego funkcjonowania rządu, podstawą do podejmowania decyzji w przemyśle i usługach, głównym składnikiem nowoczesnych programów na wszystkich poziomach nauczania.

Przyszłość statystyki bez szeroko zakrojonego działania prospołecznego, tzn. bez wychodzenia na zewnątrz i doradztwa, może być problematyczna. Wychodzenie na zewnątrz i doradztwo muszą stać się podstawowymi składnikami naszego planu strategicznego.

Kwestia ta jest podobna do tego, co widzimy w innych dziedzinach nauki. Jest ona odzwierciedleniem napięcia pomiędzy wąsko rozumianą pogonią za nauką a podejściem, które jest bardziej holistyczne (całościowe), kładzie nacisk na syntezę i zachęca do interdyscyplinarności.

Pogląd inkluzyjny promuje statystyków o szerokich horyzontach, którzy nie są przywiązani do wąskich specjalności. Zamiast tego widzą wolny świat, holistycznie włączając w to świat statystyki. Są raczej wykwalifikowani w poruszaniu się w szybko zmieniających się otoczeniach. Wiedzą jak rozwiązywać rzeczywiste problemy i mają realne wpływy.

Statystycy mogą pracować w wielu gałęziach nauki, posiadając bogatą wiedzę o aspektach ich własnej dyscypliny. W świecie narastających specjalizacji należy połączyć wiedzę generalistów z umiejętnościami specjalistów. To połączenie jest rzadkie, ale jest na nie duże zapotrzebowanie i daje ono ciekawą pracę na całe życie.

Wydaje się, że inkluzyjne spojrzenie na statystykę i wszystko co ono ze sobą niesie dla badań i karier jest sposobem, w jaki powinniśmy określić przyszłość statystyki.

Metodom statystycznym przypisuje się potrójną rolę: opisu, analizy i przewidywania (prognozowania). Podstawowym procesem przy opracowywaniu wyników badań jest analiza, która może występować w formie:

- przedmiotowej — ocena uzyskanych wyników z punktu widzenia przedmiotu badania i wyprowadzenia praktycznych wniosków wynikających z wykrytych prawidłowości,
- metodologicznej — ocena otrzymanych wyników z punktu widzenia ich merytoryczności, dokładności i wiarygodności.

Interdyscyplinarny charakter statystyki wymaga dla jej poznania znajomości podstawowej wiedzy z matematyki i informatyki, w zakresie:

- matematyki: algebry zbiorów, funkcji jednej i wielu zmiennych, elementów rachunku różniczkowego i całkowego, wybranych zagadnień kombinatoryki, rachunku wektorowego i macierzowego oraz elementów geometrii analitycznej,

- informatyki: kodowania informacji, struktury danych alfanumerycznych i systemów informacyjnych, sposobów transmisji danych, zakładania i korzystania z baz danych, stosowania standardowych komputerowych pakietów statystycznych i arkusza kalkulacyjnego oraz wyszukiwania danych na portalach internetowych.

Statystyka, przy wykorzystaniu własnych, uznanych za uniwersalne, metod i narzędzi, stara się wyjaśniać oraz umożliwiać interpretację powstałych i wyróżnionych w badaniu faktów. W ten sposób wyraża się wzajemna relacja między statystyką oraz każdą dyscypliną naukową stosującą metody statystyczne. Odnosi się to szczególnie do nauk:

- przyrodniczych (np. biologia, ekologia, fizyka, chemia, astronomia),
- ekonomicznych (np. ekonomia, finanse, organizacja, zarządzanie, marketing),
- społecznych (np. socjologia, psychologia, pedagogika),
- technicznych (np. inżynieria, technika, logistyka) itp.

Coraz mocniejsze przenikanie statystyki do tych dziedzin wiedzy doprowadziło do wyodrębnienia się różnych gałęzi nauki, takich jak: biometria, demografia, ekonometria, fizyka statystyczna, termodynamika statystyczna, psychologia i socjologia statystyczna czy też statystyka gospodarcza.

DUŻE BAZY — BIG DATA

Statystyka z całą pewnością stymuluje rozwój gospodarczy naszego kraju. Powinniśmy zwracać uwagę na współpracę w ramach szeroko rozumianych innych dyscyplin opartych na statystyce. Warto wskazać na dziedziny takiej współpracy, która staje się potrzebą dnia, a mianowicie:

- *data mining*,
- techniki wizualizacyjne.

Data mining, czasami nazywane jest „odkrywaniem wiedzy w bazach danych” (*Knowledge Discovery In Databases*), co ściśle wiąże się z *machine learning* (dość dobrze zdefiniowany przedmiot). Wikipedia, darmowa internetowa encyklopedia, podaje następujące definicje *data mining* i *machine learning*:

- *Data mining jest to znaczące wydobywanie ukrytych i wcześniej nieznanymi, a jednocześnie potencjalnie przydatnych informacji z dużych zbiorów danych lub baz danych.*
- *Machine learning jest to metoda tworzenia programów komputerowych przy wykorzystaniu analizy zbiorów danych, a nie tylko intuicji inżynierów.*

Szupiluk (2013) podaje szerszą definicję: *Data mining jest dyscypliną wyrosłą na styku problemów biznesowych, bazodanowych i analitycznych technologii informatycznych oraz metod analizy danych, także (a może przede wszystkim) tych związanych z pojęciami sztucznej inteligencji, sieci neuronowych oraz uczenia maszynowego.*

Te definicje sugerują konieczność interdyscyplinarnego szkolenia w zakresie statystyki, techniki baz danych oraz informatyki. Łatwo zauważyć, że w większości przypadków technika wykorzystywana przy *data mining* wywodzi się z techniki statystycznej, a większość narzędzi używanych przy *data mining* to te same narzędzia, które zostały opracowane na potrzeby statystycznej analizy danych.

Obecnie istnieje wiele produktów powiązanych z wizualizacją. Często dodatkowo po słowie wizualizacja dodaje się różne słowa na określenie rozmaitych rodzajów wizualizacji, np. wizualizacja danych, wizualizacja naukowa, wizualizacja informacji. W przeciwieństwie do *data mining*, wizualizacja nie jest dobrze zdefiniowanym pojęciem. Jednakże podstawową zasadę leżącą u podstaw wizualizacji można ująć następująco: ludzie myślą wizualnie, podczas gdy komputer pracuje analitycznie.

Według G. Scotta Owena *wizualizacja jest w gruncie rzeczy procesem odwzorowania — od obrazu komputerowego do obrazu percepcyjnego, przy użyciu technik kodujących w celu maksymalizacji ludzkiego zrozumienia i komunikacji*. Dlatego jest to forma eksploracyjnej analizy danych sformułowana przez Johna Turkeya. To co jest najbardziej wartościowe, jeżeli chodzi o technologie w *data mining* i wizualizacji, pochodzi głównie ze statystyki, stąd nasze wspólne pole do dyskusji i rozwoju.

Trendy w wyzwaniach stojących przed nami są interesującymi wskaźnikami rozwoju statystyki na początku XXI w. Można je scharakteryzować następująco:

- 1) zaskakujący jest trend w kierunku rozwiązywania małych problemów. W przemyśle, handlu i usługach naciski konkurencji są tak silne, horyzonty czasowe tak krótkie, zaś zbieranie danych tak kosztowne, że jesteśmy zmuszeni do niewygodnych sytuacji, w których plany i decyzje muszą być podejmowane na podstawie coraz mniejszej ilości danych;
- 2) oczywiście, jest nacisk zewnętrzny dotyczący rozwiązywania coraz większych problemów. Statystycy, dobrze wykształceni i mający umiejętność przystosowywania się do zmian, mają tendencję do stymulowania tego kierunku;
- 3) co więcej, w wielu zastosowaniach widzimy nieprawdopodobne ilości informacji, np. gigabajty w ruchu telefonicznym czy terabajty danych o globalnych zmianach klimatycznych.

Sam rozmiar nie jest najważniejszą kwestią. Jest nią kombinacja rozmiaru i złożoności — ona jest wyzwaniem. Te tzw. problemy danych masowych często nie nadają się do zaadaptowania do standardowych rozwiązań statystycznych.

Gdy statystycy ledwie zaczęli zmagać się z wyzwaniami stawianymi przez masowe zbiory danych, potencjalne, wartościowe podejścia już się wyłaniają, jak np.:

- losowanie adaptacyjne (ucząc się w trakcie),
- wizualizacja kierunkowa,
- poleganie na aproksymacji (zapomnij o optymalizacji),

- praca rozdzielona (ludzie, maszyny),
- dziel i zdobywaj (konsoliduj później),
- eksploatuj kontekst („problem zerowy”).

ZASTOSOWANIE METOD STATYSTYCZNYCH W ARCHIWIZACJI DUŻYCH ZBIORÓW DANYCH

W 2001 r. analitycy rynkowi, współpracujący z Gartner Inc., zdefiniowali *Big Data* jako zbiór trzech *V*:

Volume (pojemność, objętość) — istnieje wiele czynników, które przyczyniają się do zwiększenia ilości danych zwłaszcza w kontekście szybkiego rozwoju narzędzi informatycznych umożliwiających ich przechowywanie przy często bardzo ograniczonym koszcie.

Velocity (prędkość) — dane przesyłane są z ogromną prędkością i muszą być odpowiednio rozpoznawane i analizowane w odpowiednim czasie, bardzo często w czasie rzeczywistym. Reakcja na pojawiające się impulsy danych stanowi wyzwanie dla większości współczesnych organizacji.

Variety (różnorodność) — obecnie istnieje ogromne zróżnicowanie w zakresie formatu dostępnych danych: strukturyzowane, w formie tradycyjnych baz, w formie plików multimedialnych, dokumentów tekstowych, e-maili itd. Odpowiednie zarządzanie tak zróżnicowanymi formatami danych wymaga wyspecjalizowanych i efektywnie funkcjonujących narzędzi informatycznych.

Analitycy SAS Institute uzupełnili zbiór *3V* o dwa elementy:

Variability (zmiennność) — prędkość przepływu danych oraz ich różnorodny charakter są silnie powiązane z wysokim poziomem ich zmienności, szczególnie w kontekście danych bardzo wysokiej częstotliwości.

Complexity (złożoność) — różnorodne źródła pochodzenia danych powodują, że właściwa analiza wymaga uporządkowania ich wewnętrznej struktury poprzez odpowiednie przekształcenia, selekcję oraz oczyszczanie wejściowego zbioru w różnorodnych systemach.

Podsumowanie

Obecny etap rozwoju społeczeństwa charakteryzuje się gwałtownym rozpowszechnieniem technologii informacyjnych. Rozwój gospodarczy zależy w decydującym stopniu od umiejętności zarządzania informacją oraz od wykorzystania zdobyczy nowoczesnych technologii. Podstawową rolę w zbieraniu, analizie, interpretacji i udostępnianiu informacji odgrywa statystyka. Fakt ten nie jest jednak w wystarczającym stopniu doceniany ani przez sfery rządzące, ani przez działaczy gospodarczych.

Dla podniesienia rangi statystyki i przekonania do jej znaczenia szerszych kręgów społecznych, zwłaszcza młodzieży, nie wystarczą jednak najbardziej na-

wet przekonujące oświadczenia samych statystyków. Uznanie wysokiej roli statystyki zależy od bardziej aktywnego podejmowania przez służby statystyczne takich tematów badań, które mogą przyczynić się do rozwiązania najważniejszych, stojących obecnie przed narodami, problemów gospodarczych i społecznych.

Do niezwykle ważnych i aktualnych tematów badań, na wyniki których oczekują użytkownicy, należą finanse międzynarodowe. Chodzi tu zwłaszcza o bardziej szczegółowe dane dotyczące bezpośrednich inwestycji zagranicznych, długu publicznego i w ogóle kondycji sektora finansowego, a także bardziej szczegółowe dane charakteryzujące dynamikę i strukturę przepływów kapitału międzynarodowego. Analiza danych statystycznych z tej dziedziny mogłaby posłużyć podejmowaniu skuteczniejszych kroków zapobiegających kryzysowi finansowemu w skali międzynarodowej.

W wieloletnim rozwoju historycznym statystyka wypracowała wiele narzędzi, które można wykorzystać do rozwiązywania licznych zadań związanych z doskonaleniem działalności przedsiębiorstw. Wielu z tych narzędzi w praktyce się nie wykorzystuje. W przedsiębiorstwach statystyka powinna wspomagać kierownictwo w realizacji zadań na wszystkich poziomach zarządzania: strategicznym, zarządczym i operacyjnym (wykonawczym). Na poziomie strategicznym najważniejszą rolę odgrywa myślenie statystyczne, polegające na umiejętności kojarzenia różnych zjawisk, na podejmowaniu decyzji opartych na: bazie informacyjnej, zrozumieniu pojęcia zmienności oraz systematyczności w działaniu.

Na poziomie zarządczym, odpowiedzialnym za wdrożenie dyrektyw podejmowanych przez strategów, myślenie statystyczne oraz stosowanie narzędzi statystycznych są niezbędne do zapewnienia właściwego projektowania wyrobów, kontroli i doskonalenia procesów oraz szkolenia.

Na poziomie operacyjnym niezbędna jest wiedza dotycząca stosowania narzędzi statystycznych: opracowanie i analiza wykresów i innych metod graficznych, planowanie doświadczeń, analiza pomiarów i analiza regresji.

Niezależnie od wykorzystania metod statystycznych w usprawnianiu bieżącej działalności, kierownictwa przedsiębiorstw powinny potrafić docenić rolę statystyki w podnoszeniu poczucia zadowolenia ich klientów oraz w pomiarze wydajności działalności przedsiębiorstw.

Obok tradycyjnej roli statystyków w biznesie, polegającej na konsultowaniu projektów oraz szkoleniu grup pracowników, istnieje potrzeba poszerzenia tej roli na szkolenie wszystkich pracowników organizacji gospodarczych. To wymaga oczywiście poszerzenia wiedzy i doświadczenia samych statystyków, a także wzbogacenia programów studiów uniwersyteckich. Wymaga to także rozszerzenia współpracy szkół wyższych z biznesem.

LITERATURA

- Domański Cz. (2011), *Statystyka nauką dla wszystkich, statystyka w służbie publicznej. Wyzwania XXI w.*, Kraków
- Szupiluk R. (2013), *Dekompozycje wielowymiarowe w agregacji predykcyjnych modeli data mining*, Warszawa

SUMMARY

The author focuses on the tasks of statistics, as a science which serves the knowledge of the world and better and safer functioning of a human. Among the challenges the author mentioned an analysis of the large sized databases and their archiving, as well as the development of the functional data analysis (FDA).

РЕЗЮМЕ

Автор статьи посвящает свое внимание задачам статистики, области науки, которая позволяет знакомиться с миром, а также лучше и безопаснее жить в нем человеку. Среди задач стоящих перед статистикой автор перечислил анализ баз данных больших размеров и их архивирование, а также развитие функционального анализа данных (FDA).