



WPROWADZENIE DO ZASTOSOWAŃ METOD STATYSTYCZNYCH I TECHNIK DATA MINING W BADANIACH NAUKOWYCH

prof. dr hab. Andrzej Sokołowski¹

Statystyka to nauka o metodach badania prawidłowości występujących w zjawiskach masowych. Taką definicję statystyki jako nauki możemy znaleźć w większości podręczników. Wydaje się więc, że statystyka nastawiona jest przede wszystkim na analizę rzeczywistości, na poszukiwanie prawidłowości poprzez analizę powtarzalnych – podobnych, ale nie identycznych – zdarzeń. Z drugiej strony – uważa się statystykę za część matematyki, nauki, której esencją są rozważania teoretyczne, dowodzenie twierdzeń wyprowadzanych logicznie z ustalonych założeń. Można więc powiedzieć, że jest statystyka teoretyczna i statystyka praktyczna. Statystyka teoretyczna rozważa prawidłowości, jakie tworzą zjawiska rozpatrywane w masie, oraz prawidłowości rządzące kształtowaniem się charakterystyk liczbowych zjawisk. Na tej podstawie proponowane są konkretne metody analityczne. Wykorzystuje je statystyka praktyczna do rozwiązywania problemów stawianych przez nauki „merytoryczne”.

Gdy spoglądamy na świat okiem statystyka, to powinniśmy we wszystkich zjawiskach starać się dostrzec oddziaływanie dwojakiego rodzaju przyczyn – przyczyn głównych i przyczyn ubocznych. Przyczyny główne wpływają w sposób trwały, ukierunkowany i jednakowy na wszystkie elementy badanej zbiorowości. One powodują powstawanie prawidłowości. Przyczyny uboczne, czyli losowe, działają różnokierunkowo i różnorodnie na elementy zbiorowości. Powodują one odchylenia od prawidłowości. Efekt oddziaływania przyczyn głównych nazywany jest składnikiem systematycznym, a efekt oddziaływania przyczyn ubocznych – składnikiem losowym. Analizując zjawisko, statystyk chce dostrzec, wyodrębnić, zmierzyć i opisać obydwie składniki. Często większym wyzwaniem jest trafny opis składnika losowego. Rozróżnienie obydwu składników i ich wyodrębnienie jest możliwe dzięki działaniu prawa (praw) wielkich liczb. Ogólna idea tego prawa da się podsumować w stwierdzeniu, że w miarę wzrostu liczby obserwowanych zdarzeń danego typu, efekty oddziaływania przyczyn ubocznych wzajemnie się znoszą, natomiast uwidaczniają się efekty oddziaływania przyczyn głównych. Nawet osoby nieznające statystyki stosują w praktyce zasadę, że średnia z kilku pomiarów daje wynik „bardziej prawdziwy” niż pomiary indywidualne.

¹ Akademia Ekonomiczna w Krakowie, Katedra Statystyki.



Jeżeli statystyk ma za zadanie poddać analizie jakieś zjawisko, to postrzega je jako zbiorowość ogólną (generalną), zbiorowość wszystkich realizacji zjawiska. Zbiorowość taka może być pewną konstrukcją teoretyczną lub zbiorowością istniejącą rzeczywiście. Jeżeli analizujemy dzienne obroty punktu sprzedaży, to zbiór wszystkich dni działania tego punktu w przeszłości i przyszłości – co może być zaskoczeniem – powinien być rozpatrywany nie tylko jako zbiorowość konkretnych dni, ale jako zbiór wszystkich możliwych poniedziałków, wtorków itd. Gdy jednak czytamy wyniki statystycznych badań preferencji wyborczych Polaków (czyli odpowiedzi na pytanie, na kogo głosowałbyś, gdyby wybory były w najbliższą niedzielę), to badana zbiorowość jest bardzo konkretna – jest to zbiór wszystkich Polaków posiadających czynne prawo wyborcze.

Na badaną zbiorowość składają się obiekty (jednostki statystyczne). Obiekty te charakteryzowane są cechami statystycznymi. Tradycyjnie cechy te, czyli właściwości jednostek statystycznych, dzieliło się na jakościowe, czyli niemierzalne, i ilościowe, czyli mierzalne. Z kolei cechy mierzalne mogą mieć skończony lub przeliczalny zbiór możliwych wartości i wtedy określamy je mianem skokowych, lub mogą przyjąć każdą wartość z określonego przedziału liczbowego (nawet nieskończonego) i wtedy nazywane są zmiennymi ciągłymi. Prowadząc jakiegokolwiek badania naukowe, musimy przede wszystkim wiedzieć, co badamy, a więc co jest naszą zbiorowością, jakie jednostki statystyczne je tworzą i jakie cechy statystyczne tych jednostek chcemy badać. Precyzyjne zdefiniowanie jednostek (badanych obiektów), zbiorowości i analizowanych cech jest warunkiem koniecznym prawidłowego stosowania metod statystycznych. Kolejnym warunkiem prawidłowego wyboru metody jest uświadomienie sobie, w jakich skalach mierzone są analizowane cechy statystyczne. Skala nominalna pozwala jedynie na identyfikację obiektów, na określenie, jaki wariant cechy jakościowej zrealizował się w konkretnym obiekcie. Przy pomiarach w skali nominalnej możemy poznać strukturę zbiorowości ze względu na cechę jakościową. Jeżeli porównujemy dwa obiekty, to w skali nominalnej możemy tylko stwierdzić, czy są one identyczne czy różne. Skala porządkowa umożliwia przyporządkowanie obiektów do uporządkowanych klas lub znalezienie liniowego uporządkowania obiektów w zależności od natężenia badanej cechy. Porównując dwa obiekty, możemy stwierdzić, który z nich jest „lepszy” w zakresie konkretnej cechy, ewentualnie uznać, że obiekty są identyczne. Wielu badaczy zapomina, że rangi określają tylko porządek obiektów, a nie są miarą natężenia cechy, i działania arytmetyczne wykonywane na rangach (a więc na przykład ich dodawanie) jest „nielegalne”. Określenie różnicy między dwoma obiektami w zakresie badanej cechy umożliwia dopiero skala przedziałowa. Tu pojawiają się liczby, które można dodawać lub odejmować. Wreszcie skala najmocniejsza, czyli skala ilorazowa, umożliwia ustalenie, „ile razy” natężenie cechy w jednym obiekcie jest większe (mniejsze) niż w drugim obiekcie. Dozwolone są tu więc wszystkie działania arytmetyczne, a punkt zerowy skali ma charakter obiektywny.

Statystyka jest w pewnym sensie nauką służebną. Proponuje metody, które umożliwiają przedstawicielom różnych dyscyplin naukowych rozwiązywanie własnych problemów naukowych. Podstawową metodologią statystyczną jest *wnioskowanie statystyczne*, czyli wnioskowanie o zbiorowości generalnej na podstawie zbiorowości próbnej.



Najpierw jednak trzeba dysponować modelami, które mogą opisywać badaną zbiorowość. Takim bardzo ogólnym modelem, który może być zastosowany do opisu zachowania się cechy w populacji, jest tzw. *zmienna losowa*. Jest to wielkość, która w wyniku „doświadczenia” przyjmuje różne wartości, przy czym przed doświadczeniem nie jesteśmy w stanie określić z absolutną pewnością, jaka wartość właśnie się pojawi (zrealizuje). Jesteśmy w stanie co najwyżej podać zbiór możliwych wartości, jakie mogą się pojawić, oraz odpowiadające im prawdopodobieństwa. Prawdopodobieństwa te muszą sumować się do jedności. Funkcja, która opisuje sposób przyporządkowania prawdopodobieństw poszczególnym wartościom (lub zbiorom wartości) zmiennej losowej, nazywa się *rozkładem prawdopodobieństwa*. Zmienne losowe są wykorzystywane do opisywania zachowania się cech w populacji i, podobnie jak cechy statystyczne, dzielą się na skokowe i ciągłe. Rozkład prawdopodobieństwa może być przedstawiany przy użyciu różnych funkcji. Najbardziej uniwersalna jest *dystrybuanta*, która podaje prawdopodobieństwo tego, że zmienna losowa przyjmie wartość mniejszą od zadanej liczby. Przy zmiennych losowych skokowych korzystamy z *funkcji rozkładu prawdopodobieństwa*, która przyporządkowuje prawdopodobieństwo konkretnym wartościom. Przy zmiennej losowej skokowej ważnym pojęciem jest „sukces”. Możemy przykładowo liczyć szanse uzyskania konkretnej liczby sukcesów w zadanej liczbie niezależnych prób (rozkład dwumianowy), w próbach zależnych (rozkład hipergeometryczny), szansę uzyskania pierwszego sukcesu w konkretnej próbie (rozkład geometryczny) lub drugiego, trzeciego itd. sukcesu (rozkład wielomianowy).

Przy zmiennych losowych ciągłych wykorzystujemy *funkcję gęstości prawdopodobieństwa*, która pokazuje, jak prawdopodobieństwo rozkłada się na przedziale zmienności wartości danej zmiennej losowej. Wśród rozkładów prawdopodobieństwa zmiennej losowej ciągłej niewątpliwie centralne miejsce zajmuje *rozkład normalny*. Funkcja gęstości tego rozkładu ma kształt dzwonowaty, symetryczny. Większość zjawisk kształtowanych przez naturę czy zakłóceń czysto losowych rozkłada się według tego prawidła. Dodatkowo wiele zmiennych losowych po prostych przekształceniach da się sprowadzić do rozkładu normalnego, a dla innych - rozkład normalny służy jako aproksymacja, czyli przybliżenie rozkładu dokładnego.

Zmienna losowa może być również opisana przy pomocy pewnych charakterystyk liczbowych, z których wiele jest jednocześnie parametrami funkcji opisujących rozkład prawdopodobieństwa. Parametr to liczba opisująca zmienną losową w sposób skrótowy, sumaryczny. Najczęściej wykorzystywane parametry to miary położenia (wartość przeciętna, modalna, mediana, kwantyle) oraz miary zmienności (wariancja, odchylenie standardowe, współczynnik zmienności, odchylenie ćwiartkowe). Kształt rozkładu jest charakteryzowany przez miary asymetrii, spłaszczenia i koncentracji.

W niektórych zagadnieniach naukowych da się teoretycznie uzasadnić występowanie w populacji konkretnego rozkładu lub konkretnej wartości parametru. Przykładem może być tu iloraz inteligencji (IQ), który jest tak skonstruowany, że jego wartość przeciętna jest równa 100.



O wiele częściej jest tak, że nie znamy ani typu rozkładu, ani wartości parametrów. I wtedy przychodzi z pomocą wspomniane wcześniej wnioskowanie statystyczne. Wnioskujemy o zbiorowości (populacji) na podstawie próby. Poprawność wnioskowania zależy przede wszystkim od tego, czy próba dobrze reprezentuje analizowaną populację, czy struktura próby jest jak najbardziej zbliżona do struktury populacji. Reprezentatywność próby jest zapewniona, jeżeli próba jest losowa. Losowość próby nie zawsze jest oczywista. Rzadko mamy możliwość wykonania losowania prostego, w którym zadbamy, aby każdy element populacji miał taką samą szansę (prawdopodobieństwo) znalezienia się w próbie. Klienci, pacjenci, interesanci, widzowie itp. „zgłaszają się” sami i my mamy możliwość tylko sprawdzenia, czy ten proces ma charakter losowy czy nie. Jeszcze trudniejsza jest sytuacja przy analizie danych demograficznych i ekonomicznych. Niektórzy badacze wręcz negują trafność rozróżniania próby i populacji. Stopa bezrobocia w województwach Polski na koniec danego roku stanowi jedyną informację, jaką w tej mierze można uzyskać. Ale przecież uzyskane wartości powstały jako efekt oddziaływania przyczyn głównych oraz losowych i tu właśnie zapewniony jest ten element losowości. Nie jest to losowanie w prostym znaczeniu tego słowa. Konkretnie wartości charakterystyk ekonomicznych, jakie obserwujemy w odniesieniu do danych obiektów ekonomicznych w danym czasie, są przykładem realizacji pewnych *mechanizmów* ekonomicznych – i te mechanizmy mogą być traktowane jako populacja (zbiór możliwych scenariuszy, mniej lub bardziej prawdopodobnych).

Wnioskowanie statystyczne obejmuje dwie grupy metod: estymacje i weryfikację hipotez statystycznych. Estymacja, czyli szacowanie, to „odgadywanie” rozkładu lub wartości parametrów w populacji na podstawie próby. Estymacja rozkładu to estymacja nieparametryczna. Najprostszą metodą jest tu obliczanie częstości oraz rysowanie histogramu, który pozwala wstępnie ocenić typ rozkładu. Estymacja parametryczna wykorzystuje pewne charakterystyki liczbowe wyliczane z próby. Są to *estymatory*, które zależą od wartości parametru populacji oraz od wyników z próby. Ponieważ próba jest losowa, to i estymator jest zmienną losową posiadającą własny rozkład prawdopodobieństwa. Wymaga się, aby estymatory były zgodne (czyli w miarę wzrostu liczebności próby coraz precyzyjniej „odgadywały” szacowany parametr), nieobciążone (średnio „trafiające” w nieznaną wartość), efektywne (zapewniające mały błąd estymacji) oraz odporne (mało wrażliwe na błędy w danych). Jeżeli przyjmujemy, że nieznaną wartość parametru jest równa ocenie (wartości estymatora) otrzymanej w próbie, to mamy do czynienia z *estymacją punktową*. Można jednakże wykorzystywać informacje o rozkładzie estymatora i konstruować tzw. *przedziały ufności*, czyli przedziały liczbowe, o których z dużą ufnością (zazwyczaj 95%) możemy powiedzieć, że zawierają w sobie nieznaną, szukaną wartość parametru.

Weryfikacja hipotez statystycznych pozwala przy pomocy testu statystycznego zweryfikować hipotezę (sąd) o rozkładzie lub parametrze populacji. Test statystyczny to procedura pozwalająca odrzucić badaną hipotezę z małym ryzykiem popełnienia błędu polegającego na odrzuceniu hipotezy prawdziwej. Ryzyko to mierzone jest tzw. *poziomem istotności*, który przez większość badaczy przyjmowany jest na poziomie 0,05. Przy korzystaniu z testu statystycznego badacz musi przede wszystkim sformułować hipotezę zerową



(rozkład jest określonego typu, parametr jest równy konkretnej liczbie, parametry w dwóch populacjach są równe itp.) oraz hipotezę alternatywną, zwaną niekiedy hipotezą badawczą. Niezmiernie ważną decyzją jest wybór właściwego testu statystycznego i sprawdzenie założeń przez niego wymaganych. *Testy parametryczne* wymagają, aby rozkład badanej cechy był określonego typu (zazwyczaj normalny), a *testy nieparametryczne* wolne są od takich założeń. Na ogół jednak testy parametryczne mają większą moc, czyli „szybciej” wykrywają odstępstwo od stanu określonego w hipotezie zerowej. Podstawowym „narzędziem” wykorzystywanym w teście jest statystyka testowa, której rozkład jest znany i w związku z tym jesteśmy w stanie ocenić, które wyniki (wartości statystyki) są mało prawdopodobne przy danej hipotezie zerowej. Dawniej oceniano się to poprzez odczytywanie z tablic tzw. wartości krytycznych i porównywanie z nimi empirycznej wartości statystyki testowej. Obecnie wszystkie statystyczne pakiety komputerowe podają *wartość p*, która jest prawdopodobieństwem otrzymania wyniku bardziej przeczącego hipotezie zerowej niż ten rezultat, który właśnie otrzymaliśmy. Hipotezę zerową należy odrzucić, jeżeli wartość *p* jest mniejsza od przyjętego poziomu istotności. Jest to prosta reguła, taka sama dla wszystkich testów statystycznych. Należy tu z naciskiem podkreślić, że jeżeli wartość *p* jest większa od poziomu istotności, to nie oznacza to udowodnienia prawdziwości hipotezy zerowej. Mówimy wtedy, że nie ma podstaw do odrzucenia hipotezy zerowej, a więc, że nie udało się nam udowodnić prawdziwości naszej badawczej hipotezy alternatywnej – i tylko tyle.

Przedstawiony powyżej zarys metod wnioskowania statystycznego pozwala na właściwe stosowanie metod statystycznych w wielu zagadnieniach naukowych. Najbardziej wskazana jest tu współpraca przedstawiciela dyscypliny merytorycznej ze statystykiem. Ten pierwszy definiuje problem, stawia pytania i hipotezy w języku swej dyscypliny, a zadaniem statystyka jest zaproponowanie metod, wyjaśnienie ich założeń i ograniczeń, przeprowadzenie obliczeń i wstępna interpretacja wyników. Pomoc statystyka jest wskazana już na etapie projektowania badania oraz gromadzenia i przygotowywania danych statystycznych. Jeżeli w tym procesie badawczym celem jest zweryfikowanie teorii, przypuszczeń lub konkretnej hipotezy, to mamy do czynienia z tzw. *Confirmatory Data Analysis*, stosującą klasyczny schemat wnioskowania statystycznego. Nie zawsze jednak hipotezy badawcze mogą być precyzyjnie sformułowane. Często możliwe jest podanie pewnych przypuszczeń lub, co bardziej typowe – zadanie pytań. I tu znajduje zastosowanie *Exploratory Data Analysis*, a więc „poszukiwawcza” analiza danych. Najbardziej typowym przykładem może tu być analiza skupień, w ramach której poszukuje się wcześniej nieznaną liczbę jednorodnych podgrup, by następnie poddać je opisowi, porównywaniu i interpretacji. W pewnym sensie zagadnienie prognozowania jest problemem analizy eksploracyjnej wówczas, gdy model prognostyczny nie jest znany (nie wynika z teorii zjawiska), a zadaniem jest zbudowanie jak najbardziej trafnej prognozy. Spośród wielu możliwych podejść da się tu zastosować sieci neuronowe, które „same” poszukują najlepszego modelu prognostycznego.

I wreszcie dochodzimy do przedmiotu i metod *data mining* (zgłębiania danych). Różne definicje, najogólniej rzecz biorąc, zwracają uwagę na to, że data mining to proces znajdowania interesujących wzorców, powiązań, zmian, anomalii, ukrytych struktur w bar-



dzo dużych zbiorach danych zgromadzonych w hurtowniach danych (bez określonego celu badawczego).

W klasycznych metodach statystycznych i większości metod eksploracyjnych komputer używany jest jako narzędzie obliczeniowe na przygotowanych wcześniej zbiorach danych. W większości zagadnień te zbiory danych są małe, uporządkowane i przygotowane już z myślą o celu badania i metodach, jakie będą stosowane. Od biedy obliczenia dałoby się wykonać na ręcznym kalkulatorze. To co jest charakterystyczne dla metodyki data mining, to ogrom danych nieuporządkowanych i konieczność zastosowania komputerów. Nie przez przypadek rozwój tych metod idzie w parze z rozwojem mocy obliczeniowej komputerów.

Typowe etapy data mining to:

- ◆ *czyszczenie danych* – identyfikacja błędów, braków danych, ich ewentualna korekta lub uzupełnianie,
- ◆ *integracja danych* – różne, często niejednorodne źródła danych łączone są w jednorodny zbiór danych,
- ◆ *wybór danych* – ze zbioru danych wybieramy te cechy i obiekty, które będą wykorzystywane w analizie,
- ◆ *przekształcenia danych* – może to być agregacja, sumowanie lub inne proste transformacje zmiennych,
- ◆ *data mining* – właściwy proces poszukiwania wzorców, prawidłowości, anomalii itp.,
- ◆ *ocena „odkrytych” wzorców, grup, powiązań itp.*,
- ◆ *prezentacja wyników* – w postaci zrozumiałej dla odbiorcy.

W data mining zazwyczaj nie korzysta się z typowych procedur klasycznej statystyki. Są tu wykorzystywane metody eksploracyjne stosowane do dużych zbiorów danych. Można zadać pytanie: czy jeżeli posiadamy bardzo liczny zbiór danych, to czy konieczna jest jego analiza całkowita (wyczerpująca) poprzez data mining, czy nie lepiej jest wylosować z tego zbioru danych próby i powrócić do schematu klasycznego wnioskowania statystycznego. Takie podejście możliwe jest wtedy, gdy ustalony jest jasno cel badań i hipotezy robocze.

Wydaje się, że wzajemne relacje pomiędzy wspomnianymi tu metodologiami można podsumować w poniższej tabeli.



	Dane	Cel badań	Wykorzystanie komputerów
WNOSKOWANIE STATYSTYCZNE	dane uporządkowane, raczej mało obiektów, mało cech	estymacja, weryfikacja hipotez	możliwe
ANALIZA DANYCH	dane uporządkowane, dużo cech, może być dużo obiektów	odpowiedź na ogólne pytania, poszukiwanie prawidłowości	pożądane
DATA MINING	dane nieuporządkowane, olbrzymia ilość danych, dużo cech i obiektów	poszukiwanie prawidłowości, wzorców, związków i anomalii	konieczne