



## PLANOWANIE I ANALIZA EKSPERYMENTÓW A TECHNIKI DATA MINING W BADANIACH EMPIRYCZNYCH. PRZYKŁADY ANALIZ W *STATISTICA*

**Janusz Wątroba**  
*StatSoft Polska Sp. z o. o.*

Problematyka sposobu wykorzystania metod analizy danych przy opracowywaniu wyników badań empirycznych ma duże znaczenie zarówno w badaniach naukowych, jak i w zastosowaniach praktycznych. W artykule zwrócono uwagę na najważniejsze różnice występujące pomiędzy dwoma powszechnie stosowanymi podejściami oraz zilustrowano je praktycznymi przykładami analiz w środowisku programu *STATISTICA*.

### Wprowadzenie

We współczesnych zastosowaniach metod analizy danych do opracowywania wyników badań empirycznych bardzo często wyróżnia się dwa umowne podejścia. Pierwsze z nich to tzw. **konfirmacyjna (potwierdzająca) analiza danych**. W podejściu tym wybór metod analizy wynika ze sposobu sformułowania problemu badawczego i postawionych hipotez lub pytań badawczych. Badacz zwykle opiera się na istniejącej lub postulowanej teorii bądź też na wynikach wcześniejszych badań. Z metodologicznego punktu widzenia jest to podejście bardziej poprawne. W drugim ze stosowanych powszechnie podejść problemy badawcze są stawiane w sposób bardziej ogólny i na ogół przyjmują postać pytań badawczych; przy zbieraniu danych wykorzystuje się zazwyczaj metodę obserwacji. Podejście to określa się terminem **eksploracyjna (poszukiwawcza) analiza danych**.

Spróbujmy przyrzeć się bliżej najważniejszym różnicom charakteryzującym te dwa podejścia. Wydaje się, że podstawowa różnica wynika bezpośrednio z celów, jakie badacz stawia sobie w procesie badawczym. W podejściu konfirmacyjnym celem jest zazwyczaj zweryfikowanie istniejącej lub postulowanej teorii bądź postawionych a priori szczegółowych hipotez. Z kolei jeśli problemy badawcze są stawiane w postaci pytań lub pewnych ogólnych przypuszczeń, wówczas stosowane jest raczej drugie z wymienionych podejść. Kolejna różnica dotyczy sposobu zbierania danych empirycznych. W pierwszym z omawianych podejść zwraca się baczną uwagę na dokładne zdefiniowanie zbiorowości obiektów, których mają dotyczyć wnioski z badań, i określenie sposobu wyboru reprezentacyjnej próby. Dane empiryczne są uzyskiwane w wyniku przeprowadzenia ściśle



zaplanowanych eksperymentów. Przy analizie zebranych danych stosowany jest klasyczny schemat wnioskowania statystycznego. W podejściu eksploracyjnym badacz zazwyczaj nie ingeruje w przebieg badanych zjawisk i procesów i zbiera dane za pomocą obserwacji. Zdarza się też, że do analizy są wykorzystywane dane, które były gromadzone dla innych celów. Różna jest też rola danych empirycznych w obu podejściach. W podejściu confirmacyjnym dane empiryczne służą do potwierdzania poprawności przyjętych a priori modeli, natomiast w przypadku podejścia eksploracyjnego dane empiryczne stanowią punkt wyjścia przy poszukiwaniu w obrębie zebranych danych prawidłowości, powiązań i wzorców.

Jedną z konsekwencji różnic pomiędzy tymi podejściami jest to, że w podejściu confirmacyjnym wykorzystywane są niemal wyłącznie klasyczne techniki wnioskowania statystycznego (np. metody weryfikacji hipotez statystycznych). Z kolei w przypadku podejścia eksploracyjnego stosowane są różne techniki graficzne, analiza w grupach przekrojowych, metody klasyfikacyjne, techniki *data mining* czy też metody uczenia maszynowego (*machine learning*). Inną niezwykle ważną konsekwencją jest różny zakres wnioskowania. W podejściu confirmacyjnym badacz ma prawo uogólnić wnioski z analizy na populację, natomiast w przypadku podejścia eksploracyjnego możliwość uogólnienia wyników badań jest zazwyczaj dość ograniczona.

Trzeba w tym miejscu zwrócić uwagę na fakt, że w wielu obszarach badań empirycznych podejście confirmacyjne jest bardzo trudne albo wręcz niemożliwe do zastosowania. Przykładowo w niektórych zagadnieniach medycznych, finansowych czy ekonomicznych trudno wyobrazić sobie możliwość przeprowadzania klasycznych eksperymentów. W wielu dziedzinach trudności mogą z kolei wynikać z występowania dużej liczby czynników (przyczyn głównych) i ich interakcji, co w zdecydowany sposób zwiększa liczbę obiektów potrzebnych do przebadania. W takiej sytuacji badacz jest niejako skazany na zastosowanie podejścia eksploracyjnego. Czasami ze względu na koszty lub ograniczenia czasowe badaczowi nie zostaje nic innego jak wykorzystanie jakościowo „gorszych” danych, które pierwotnie były gromadzone dla celów sprawozdawczości lub przy okazji przeprowadzania badań przesiewowych.

Omawiane podejścia niekiedy wykorzystują te same techniki analizy danych. Dobrym przykładem, trafnie ilustrującym to stwierdzenie jest jedna z najbardziej popularnych metod analizy danych, jaką jest analiza regresji. Przy podejściu confirmacyjnym z góry zakładamy funkcyjną postać zależności pomiędzy zmienną niezależną (objaśniającą) a zmienną zależną (objaśnianą), np. przyjmujemy liniowy model związku, następnie dla badanego zbioru obiektów ustalamy poziomy zmiennych niezależnych i rejestrujemy wartości zmiennej zależnej. Z kolei na podstawie zebranych danych szacujemy parametry modelu i sprawdzamy, czy dane empiryczne potwierdzają liniowość związku. Przykładem takiego postępowania może być badanie typu „dawka-efekt”. Analiza regresji jest też często wykorzystywana w podejściu eksploracyjnym. Jednak w tym przypadku badacz obserwuje w danej grupie obiektów wartości pewnych zmiennych, a następnie już po zebraniu empirycznych danych próbuje do nich dopasować model najlepiej odzwierciedlający zaobserwowane powiązania. Dalsze postępowanie jest podobne jak wyżej, z tą



różnicą, że brak jest wystarczających podstaw do uogólnienia tego modelu na jakąś szerszą zbiorowość obiektów.

W obydwu podejściach jest też zazwyczaj przeprowadzana tzw. wstępna analiza danych (badanie rozkładów, obliczanie statystyk opisowych, badanie liniowości związku), jednak jej cel bywa nieco odmienny. W podejściu confirmacyjnym sprawdza się, czy zebrane dane empiryczne spełniają założenia stawiane przez metody wnioskowania statystycznego, natomiast w podejściu eksploracyjnym analiza wstępna umożliwia wybór docelowych metod analizy danych.

W dalszej części artykułu omawiane podejścia zostaną zilustrowane praktycznymi przykładami analiz. W pierwszej kolejności opiszemy przykład planowania i analizy wyników eksperymentu, natomiast w drugim przykładzie pokażemy eksploracyjną analizę natężenia wybranych czynników ryzyka choroby niedokrwiennej serca oraz budowanie modeli pozwalających na przewidywanie wystąpienia choroby niedokrwiennej serca.

## Zastosowanie podejścia confirmacyjnego przy opracowywaniu wyników przykładowego eksperymentu

Bardzo dobrym przykładem confirmacyjnej analizy danych jest podejście stosowane w **planowaniu i analizie eksperymentów**. Z punktu widzenia potrzeb badacza jest to podejście szczególnie zalecane, gdyż z jednej strony zapewnia respektowanie wymogów metodologicznych, stawianych w przypadku badań empirycznych, z drugiej zaś pozwala poprawnie zaplanować badanie oraz również oferuje gotowe procedury opracowania wyników.

Dobre zaplanowanie eksperymentu nie jest sprawą łatwą i wymaga od badacza szerokiej wiedzy z zakresu uprawianej dziedziny, bogatego doświadczenia, a niekiedy także dużej pomysłowości. Cenną pomocą może również okazać się prześledzenie doświadczeń przeprowadzanych przez innych badaczy. Spośród dostępnych na ten temat publikacji warto polecić znakomitą książkę Cobba (1998). Jest to jedna z tych pozycji, w której autor rzeczywiście, zgodnie z tytułem pisze nie tylko o technikach analizy danych eksperymentalnych, ale również o samym planowaniu doświadczeń.

Mówiąc najkrócej, chodzi o to, aby zaaranżować pomiar określonych **własności** badanych **jednostek w warunkach** umożliwiających weryfikację postawionych przez badacza hipotez i pytań badawczych. Wymaga to oczywiście wcześniejszego podjęcia decyzji, jakie konkretne pomiary będą przeprowadzane, jakie jednostki zostaną poddane badaniom oraz jakie warunki będą porównywane. Efektem dobrego zaplanowania doświadczenia jest uzyskanie danych, które są następnie opracowywane za pomocą odpowiedniego modelu analizy wariancji.

Dla ilustracji planowania i analizy eksperymentu wykorzystamy przykład opisany we wspomnianej wyżej książce Cobba. Celem badań była ocena wpływu wybranych warunków zewnętrznych środowiska na proces hibernacji u zwierząt. Nadejście pory zimowej wiąże się z dwoma rodzajami zmian. Po pierwsze spada temperatura, a po drugie



dni stają się krótsze. W opisywanym eksperymencie skupiono się na tym drugim rodzaju zmian. Badane zwierzęta przebywały w pomieszczeniach o różnej długości dni i nocy w ciągu doby. Kolejna decyzja była związana z wybraniem odpowiedniej miary wpływu długości dnia. W grę mógłby wchodzić pomiar częstości skurczów serca zwierzęcia lub częstość oddechów, gdyż praca serca i płuc u zwierząt w stanie snu zimowego ulega spowolnieniu, ale zjawisko to jest powszechnie znane i z biologicznego punktu widzenia mało interesujące. Zamiast tego postanowiono zmierzyć jakiś parametr określający poziom aktywności układu nerwowego zwierzęcia. W związku z tym wybrano stężenie pewnego enzymu regulującego przebieg reakcji biochemicznej zwanej pompą sodową, która jest odpowiedzialna za transmisję impulsów nerwowych ( $\text{Na}^+\text{K}^+\text{ATP-aza}$ ). Stężenie tego enzymu jest różne w różnych częściach organizmu. Postanowiono mierzyć je w obrębie dwóch organów: serca i mózgu. Kolejna decyzja dotyczyła wyboru odpowiedniego gatunku zwierząt. Biorąc pod uwagę koszty oraz warunki techniczne, do eksperymentu zdecydowano się wybrać chomiki.

Druga część procesu planowania eksperymentu polega na wyborze sposobu przypisania warunków do badanych jednostek. W zależności od charakteru interesujących nas warunków i badanych jednostek mamy do wyboru dwie techniki: **randomizację** i **blokowanie**. Randomizacja polega na losowym doborze jednostek, w stosunku do których są stosowane odpowiednie warunki (zabiegi). Przykładowo przy badaniu skuteczności dwóch różnych leków spośród pewnej grupy pacjentów losujemy tych, którym zostanie podany lek A, oraz tych, którzy otrzymają lek B. Czasami zdarzają się sytuacje, kiedy wiemy, że istnieją czynniki mające duży wpływ na badane przez nas cechy i chcielibyśmy ten wpływ w jakiś sposób uwzględnić. Wtedy właśnie stosuje się blokowanie, które pozwala zmniejszyć zmienność spowodowaną niejednorodnością badanych obiektów. Przykładami zmiennych blokujących mogą być: wielkość dochodu, poziom wykształcenia lub doświadczenie. Niekiedy mogą to być również charakterystyki samego eksperymentu, np. czas jego trwania lub partia materiału, z którego pobierano jednostki. Przy planowaniu doświadczenia trzeba również wybrać sposób wzajemnego rozmieszczenia warunków (czynników, zabiegów). Największe możliwości wnioskowania daje tzw. **układ czynnikowy kompletny**, w którym występują wszystkie możliwe kombinacje poziomów branych pod uwagę czynników.

Opisywany eksperyment został zaplanowany bardzo efektywnie. Badaniom poddano 8 chomików. Cztery z nich (losowo wybrane) przebywały w pomieszczeniu, w którym przez 16 godzin na dobę było jasno, a przez pozostałe 8 godzin ciemno, natomiast pozostałe cztery w pomieszczeniu, w którym przez 8 godzin w ciągu doby było jasno, a przez pozostałe 16 godzin ciemno. U każdego z chomików przeprowadzono dwa pomiary stężenia  $\text{Na}^+\text{K}^+\text{ATP-azy}$ , jeden pomiar w obrębie serca i jeden pomiar w obrębie mózgu. Otrzymany układ doświadczalny nazywa się **dwuczynnikowym układem analizy wariancji**, przy czym jeden z czynników jest czynnikiem powtarzanych pomiarów. Tak zaplanowany eksperyment pozwolił sformułować rozważany problem badawczy w postaci trzech szczegółowych pytań:

- ◆ Czy długość dnia wpływa na stężenie enzymu regulującego reakcję pompy sodowej u badanych zwierząt, a jeśli tak, to jak duży jest to efekt?

- ◆ Czy w sercu i mózgu występują różne średnie stężenia badanego enzymu oraz jak duża jest ta różnica?
- ◆ Czy różnica występująca pomiędzy stężeniami enzymu w obrębie serca i w obrębie mózgu jest taka sama w przypadku różnej długości dnia i nocy w ciągu doby? (jest to pytanie o interakcję).

Śledząc kolejne kroki występujące w trakcie projektowania eksperymentu można przekonać się, że znajomość podstaw planowania i analizy doświadczeń jest nieodzownym warunkiem dobrego zaprojektowania eksperymentu. Należy jednak pamiętać, że podstawową rolę odgrywa znajomość merytorycznych zagadnień badanej dyscypliny naukowej.

Poniżej przedstawiono arkusz danych zawierający wyniki zebranych pomiarów.

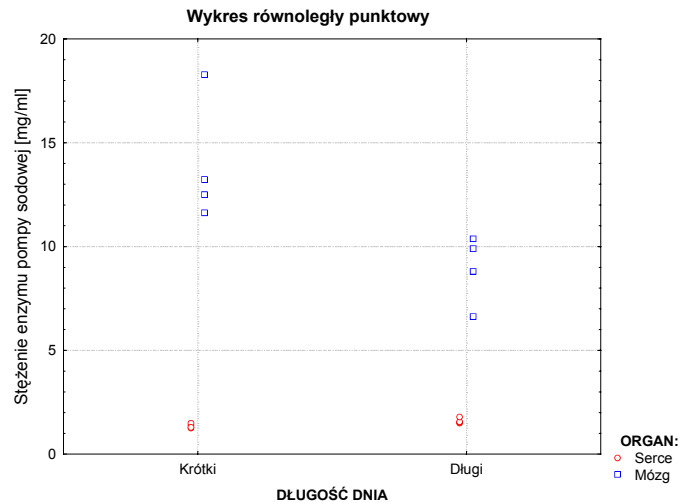
Dane zawierają informacje na temat stężenia enzymu Na <sup>+</sup> K <sup>+</sup> ATP-azy oznaczanego w sercu i mózgu badanych chomików. Źródło: Cobb G. W. (1998), <i>Design and Analysis of Experiments</i> , Springer.			
	1 Długość dnia	2 Mózg	3 Serce
1	Długi	6,625	1,490
2	Długi	10,375	1,525
3	Długi	9,900	1,555
4	Długi	8,800	1,790
5	Krótki	12,500	1,385
6	Krótki	11,625	1,485
7	Krótki	18,275	1,255
8	Krótki	13,225	1,285

Przed formalnym przeanalizowaniem wyników eksperymentu z użyciem odpowiednich procedur analizy wariancji przeprowadzimy wstępną ocenę zebranych danych. Prosty sposób podsumowania wyników przeprowadzonego eksperymentu jest zestawienie odpowiednich średnich. Dla prezentowanego przykładu średnie zostały przedstawione w poniższej tabeli.

Wyniki analizy przekrojowej (Hibernacja.sta), N=8			
Długość dnia	Mózg Średnie	Serce Średnie	Średnie grupowe
Krótki	13,906	1,353	7,629
Długi	8,925	1,590	5,258
Średnie grupowe	11,416	1,471	6,443

Średnie mają w takiej sytuacji dwie podstawowe zalety. Każda średnia ujmuje sumarycznie grupę obserwacji w postaci jednej liczby, w ten sposób pozbywamy się szczegółów, co pozwala lepiej dostrzec występujące prawidłowości. Oprócz tego średnie ułatwiają przeprowadzenie ilościowych porównań. Na podstawie wyników zawartych w tabeli możemy zauważyć, że średnie stężenie badanego enzymu obserwowane w obrębie serca chomików przebywających w pomieszczeniach o krótszym czasie trwania pory dziennej wyniosło 1,353 mg/ml i było o 0,243 mg/ml niższe od średniego stężenia dla długich dni. Z kolei średnie stężenie Na<sup>+</sup>K<sup>+</sup>ATP-azy w obrębie mózgu rejestrowane u zwierząt przebywających krócej w świetle dziennym kształtowało się na poziomie 13,906 mg/ml i było o 4,981 mg/ml wyższe od średniego stężenia rejestrowanego u chomików przetrzymywanych dłużej w świetle dziennym.

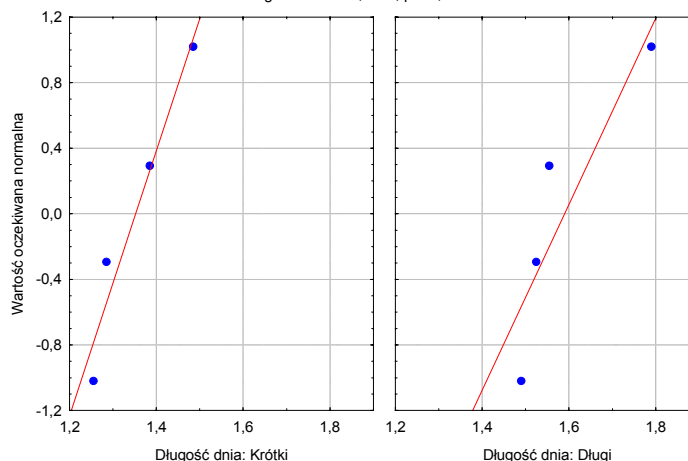
W niektórych sytuacjach stosowanie średnich może jednak nie być wskazane. Na przykład wtedy, gdy z jakichś względów ważne są szczegóły, które ukrywa operacja liczenia średniej lub zachodzi potrzeba porównania ze sobą większej liczby grup. Dobrym rozwiązaniem może wtedy okazać się tzw. wykres równoległy punktowy. Dla danych pochodzących z opisywanego eksperymentu został on przedstawiony poniżej.

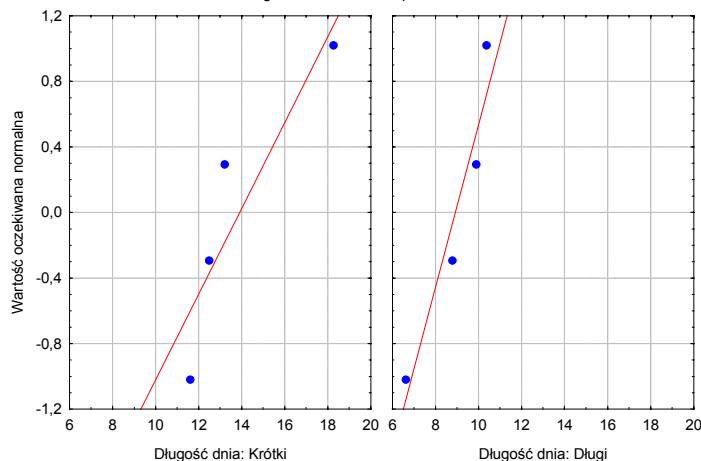


Główną zaletą takiego wykresu jest uwzględnianie wszystkich danych na jednym obrazie, co pozwala łatwo uchwycić występujące ogólne prawidłowości. Zaprezentowany wyżej wykres umożliwia stwierdzenie, że niezależnie od długości utrzymywania światła dziennego stężenie badanego enzymu było zdecydowanie niższe w obrębie serca badanych zwierząt. Ponadto wykres daje możliwość zauważenia odstających obserwacji oraz ocenę stopnia rozrzutu pojedynczych obserwacji w obrębie grup. Na przykład widać, że u jednego z chomików hodowanych w warunkach krótszego utrzymywania światła dziennego stężenie  $\text{Na}^+\text{K}^+\text{ATP}$ -azy w obrębie mózgu jest wyraźnie wyższe niż u pozostałych trzech osobników hodowanych w takich samych warunkach. Rozmieszczenie pojedynczych punktów na wykresie wskazuje też, że w obrębie mózgu badanych zwierząt rozrzut pojedynczych obserwacji jest zdecydowanie większy niż w przypadku stężeń rejestrowanych w obrębie serca.

Przed przejściem do formalnej oceny istotności poszczególnych efektów powinniśmy jeszcze sprawdzić, czy są spełnione założenia wymagane w przypadku analizy wariancji. Dwa najważniejsze dotyczą normalności rozkładu analizowanych zmiennych zależnych oraz równości wariancji w obrębie grup. Dla oceny normalności rozkładu utworzono odpowiednie wykresy normalności oraz przeprowadzono analityczny test Shapiro-Wilka. Wyniki przedstawiają zamieszczone poniżej wykresy.

**Wykres normalności dla zmiennej Serce (Hibernacja.sta)**

 Krótki: SW-W = 0,9341; p = 0,6187  
 Długi: SW-W = 0,8042; p = 0,1100

**Wykres normalności dla zmiennej Mózg (Hibernacja.sta)**

 Krótki: SW-W = 0,8232; p = 0,1507  
 Długi: SW-W = 0,9115; p = 0,4902


Z układu punktów na wykresach normalności możemy wnioskować, że założenie normalności jest spełnione (umieszczona na wykresie prosta jest wykreślana przy założeniu rozkładu normalnego). Potwierdzają to również wyniki testu Shapiro-Wilka. Z kolei przytoczone poniżej wyniki testów jednorodności wariancji upoważniają nas do stwierdzenia, że również założenie równości wariancji jest spełnione.

	Testy jednorodności wariancji (Hibernacja.sta)				
	F-max Hartleya	C Cochrana	Chi-kwadrat Bartletta	df	p
Mózg	3,198	0,762	0,824	1	0,36404
Serce	1,697	0,629	0,178	1	0,67331

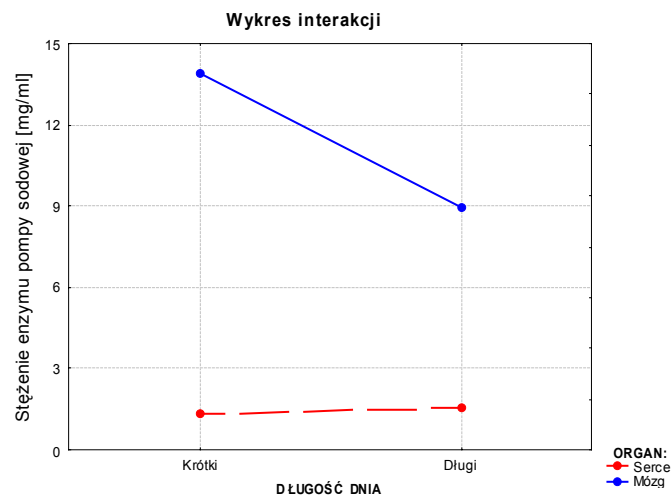
W ostatnim etapie analizy możemy ocenić statystyczną istotność efektów, aby upewnić się, że oceniane różnice nie są spowodowane specyficznym układem danych (rozkład, jednorodność), lecz mogą być uogólnione na populację. Najważniejsze wyniki przedstawia poniższa tabela analizy wariancji.

Efekt	Analiza wariancji z powtarzanymi pomiarami (Hibernacja.sta)				
	SS	df	MS	F	p
<b>Wyraz wolny</b>	<b>664,286</b>	<b>1</b>	<b>664,286</b>	<b>235,312</b>	<b>0,00000</b>
Długość dnia	22,503	1	22,503	7,971	0,03022
Błąd	16,938	6	2,823		
ORGAN	395,562	1	395,562	130,115	0,00003
ORGAN*Długość dnia	27,235	1	27,235	8,959	0,02422
Błąd	18,241	6	3,040		

Zamieszczone powyżej wyniki pozwalają na wyciągnięcie następujących wniosków:

- ◆ Długość dnia istotnie wpływa na zróżnicowanie przeciętnego poziomu badanego enzymu (u chomików hodowanych w warunkach krótszego utrzymywania światła dziennego stężenie  $\text{Na}^+\text{K}^+\text{ATP}$ -azy było przeciętnie wyższe o ponad 2 mg/ml).
- ◆ Przeciętne stężenie  $\text{Na}^+\text{K}^+\text{ATP}$ -azy zmierzone w obrębie mózgu okazało się istotnie wyższe od stężenia  $\text{Na}^+\text{K}^+\text{ATP}$ -azy w obrębie serca (różnica wyniosła blisko 10 mg/ml).
- ◆ W przypadku serca przeciętne stężenie badanego enzymu dla krótszej pory dziennej było nieco niższe od przeciętnego stężenia obserwowanego przy dłuższym utrzymywaniu pory dziennej (o około 0,25 mg/ml), natomiast w przypadku mózgu sytuacja przedstawia się odwrotnie: przeciętne stężenie  $\text{Na}^+\text{K}^+\text{ATP}$ -azy przy krótkim czasie trwania pory dziennej było zdecydowanie wyższe w porównaniu z przeciętnym stężeniem u chomików hodowanych przy dłuższym czasie trwania pory dziennej (o około 5 mg/ml).

Trzeci z przytoczonych powyżej wniosków dotyczy efektu interakcji. Dobrze ilustruje go zamieszczony poniżej wykres.





Zgodnie z planem przedstawionym na początku ostatni fragment artykułu będzie zawierał krótką charakterystykę podejścia eksploracyjnego do analizy danych empirycznych oraz praktyczny przykład analizy wykonanej z wykorzystaniem środowiska programu *STATISTICA*.

## Wybrane aspekty i przykład wykorzystania technik eksploracyjnych przy opracowywaniu danych medycznych

W obecnych czasach coraz częściej mamy do czynienia z sytuacją, gdy ilość dostępnych informacji różnego rodzaju przerasta nasze możliwości percepcyjne. Prowadzi to do coraz silniej odczuwanej potrzeby dostępu do odpowiednich narzędzi, umożliwiających modelowanie i analizę danych. Do wydobywania użytecznej wiedzy z danych są z powodzeniem wykorzystywane metody statystyki stosowanej. Od pewnego czasu uformowało się również nowe podejście do modelowania i analizy dużych ilości danych, które określa się terminem *data mining*. W literaturze przedmiotu podawanych jest wiele określeń wyjaśniających znaczenie tego terminu. Jedno z nich mówi, że *data mining* to proces selekcji, eksploracji i modelowania dużych zbiorów (baz) danych, którego celem jest odkrywanie nieznanych a priori prawidłowości, powiązań i wzorców (Giudici 2003). Z definicji tej w szczególności wynika, że podejście do analizy danych w *data mining* różni się od statystyki stosowanej głównie celem. Statystyka stosowana zajmuje się zastosowaniem metod statystycznych w odniesieniu do zgromadzonych wcześniej danych, natomiast *data mining* obejmuje cały proces wydobywania oraz analizy danych i ma zazwyczaj na celu tworzenie rozwiązań wspierających zagadnienia decyzyjne, a nie budowę lub sprawdzanie teorii. Przykładem takich zastosowań może być poszukiwanie reguł do oceny ryzyka wystąpienia określonych chorób w medycynie czy też poszukiwanie grup klientów, do których warto skierować nową ofertę promocyjną w zastosowaniach biznesowych.

W prezentowanym dalej przykładzie wykorzystano dane pochodzące z badań przeprowadzonych na terenie trzech przemysłowych rejonów w RPA (Hastie i wsp. 2001). Celem tych badań było określenie natężenia wybranych czynników ryzyka choroby niedokrwiennej serca. Badaniami objęto 462 białych mężczyzn w wieku od 15 do 64 lat. U 163 badanych stwierdzono wystąpienie zawału serca, natomiast pozostałych 299 mężczyzn stanowiło grupę porównawczą. Dla każdego z badanych zebrano informacje o wystąpieniu choroby serca oraz o czynnikach, które zwiększają ryzyko jej wystąpienia: występowanie chorób serca w rodzinie badanego, wiek, skurczowe ciśnienie krwi, poziom cholesterolu (LDL), stopień otyłości, zachowanie typu A, palenie papierosów.

Najpierw przeprowadzimy analizę niezależnego wpływu poszczególnych czynników ryzyka za pomocą analizy przekrojowej i tabel wielodzzielczych, a następnie przy budowie modeli wykorzystamy uogólnioną analizę dyskryminacyjną (GDA) oraz technikę drzew klasyfikacyjnych.

Poniżej zamieszczono fragment pliku danych.

Badanie czynników ryzyka choroby wieńcowej	Jest to fragment danych pochodzących z badań przeprowadzonych na terenie trzech przemysłowych rejonów w RPA (Rousseau i wsp. 1983). Celem badań było określenie natężenia czynników ryzyka choroby niedokrwiennej serca. Badaniami objęto białych mężczyzn w wieku od 15 do 64 lat. U 163 badanych stwierdzono wystąpienie zawału serca, natomiast pozostałych 299 mężczyzn stanowiło grupę porównawczą. Arkusz danych zawiera wyniki pomiarów wybranych czynników podwyższających ryzyko wystąpienia chorób serca.							
	1 Choroba wieńcowa	2 Wiek	3 Skurczowe ciśnienie krwi	4 Poziom cholesterolu (LDL)	5 Otyłość	6 Zachowanie typu A	7 Palenie	8 Choroby serca w rodzinie
1	Tak	52	160	5,73	23,11	49	12,00	Tak
2	Tak	63	144	4,41	28,61	55	0,01	Nie
3	Nie	46	118	3,48	32,28	52	0,08	Tak
4	Tak	58	170	6,41	38,03	51	7,50	Tak
5	Tak	49	134	3,50	27,78	60	13,60	Tak
6	Nie	45	132	6,47	36,21	62	6,20	Tak
7	Nie	38	142	3,38	16,20	59	4,05	Nie
8	Tak	58	114	4,59	14,60	62	4,08	Tak
9	Nie	29	114	3,83	19,40	49	0,00	Tak

Dla oceny zróżnicowania przeciętnego poziomu analizowanych zmiennych ilościowych (Wiek, Skurczowego ciśnienia krwi, Poziomu cholesterolu frakcji LDL, Otyłości, Zachowania typu A oraz Palenia) w obu grupach badanych mężczyzn przeprowadzono analizę przekrojową. Jej wyniki pozwalają pośrednio ocenić wpływ tych czynników na występowanie choroby wieńcowej. Wyniki analizy przedstawiono w tabeli poniżej.

Analiza przekrojowa; średnie w grupach (n=462)						
Choroba wieńcowa	Wiek	Skurczowe ciśnienie krwi	Poziom cholesterolu (LDL)	Otyłość	Zachowanie typu A	Palenie
	Średnie	Średnie	Średnie	Średnie	Średnie	Średnie
Nie	38,68	135,38	4,33	23,87	52,34	2,63
Tak	50,41	143,73	5,50	28,23	54,51	5,48
Cały zbiór	42,82	138,33	4,74	25,41	53,10	3,64

Oceniono również zróżnicowanie przeciętnego poziomu analizowanych zmiennych w badanych grupach (ze względu na niespełnienie założenia normalności rozkładów zastosowano nieparametryczny test Manna-Whitneya). Dla wszystkich zmiennych zróżnicowanie okazało się statystycznie istotne przy  $p < 0,05$ . Biorąc pod uwagę relatywne różnice, mężczyźni, u których wystąpiła choroba wieńcowa, cechowali się przeciętnie starszym wiekiem, wyższym poziomem cholesterolu oraz palili więcej papierosów. Dla pozostałych czynników różnice są mniejsze.

W badaniach stosowanych w medycynie dość powszechnym zabiegiem jest sprowadzanie zmiennych ilościowych do postaci zmiennych skategoryzowanych (np. poniżej normy i powyżej normy). W takim przypadku zazwyczaj pojawia się problem ustalenia tzw. punktu lub punktów odcięcia (cut-off points). W opisywanym przykładzie cały zakres zmienności uwzględnionych zmiennych ilościowych podzielono na dwie części. Jako punkt odcięcia przyjęto średnią lub wartość normy (np. dla poziomu cholesterolu przyjęto normę 3,5 mmol/l). Następnie dla oceny wpływu tak określonych zmiennych jakościowych na częstość występowania choroby wieńcowej utworzono odpowiednie tabele dwudzielcze.

Choroba wieńcowa	Tabela dwudzielcza		
	Choroby serca w rodzinie	Choroby serca w rodzinie	Wiersze
	Nie	Tak	Razem
Nie	206	93	299
Kolumna %	76.30%	48.44%	
Tak	64	99	163
Kolumna %	23.70%	51.56%	
Razem	270	192	462

Choroba wieńcowa	Tabela dwudzielcza		
	Wiek kat	Wiek kat	Wiersze
	Młodszy	Starszy	Razem
Nie	157	142	299
Kolumna %	83.51%	51.82%	
Tak	31	132	163
Kolumna %	16.49%	48.18%	
Razem	188	274	462

Choroba wieńcowa	Tabela dwudzielcza		
	Ciśn kat	Ciśn kat	Wiersze
	Niższe	Wyższe	Razem
Nie	212	87	299
Kolumna %	70.20%	54.37%	
Tak	90	73	163
Kolumna %	29.80%	45.63%	
Razem	302	160	462

Choroba wieńcowa	Tabela dwudzielcza		
	Chol kat	Chol kat	Wiersze
	Niższy	Wyższy	Razem
Nie	107	192	299
Kolumna %	80.45%	58.36%	
Tak	26	137	163
Kolumna %	19.55%	41.64%	
Razem	133	329	462

Choroba wieńcowa	Tabela dwudzielcza		
	Otył kat	Otył kat	Wiersze
	Mniejsza	Większa	Razem
Nie	158	141	299
Kolumna %	76.70%	55.08%	
Tak	48	115	163
Kolumna %	23.30%	44.92%	
Razem	206	256	462

Choroba wieńcowa	Tabela dwudzielcza		
	Zach A kat	Zach A kat	Wiersze
	Mniej	Więcej	Razem
Nie	159	140	299
Kolumna %	68.24%	61.14%	
Tak	74	89	163
Kolumna %	31.76%	38.86%	
Razem	233	229	462

Choroba wieńcowa	Tabela dwudzielcza		
	Pal kat	Pal kat	Wiersze
	Mniej	Więcej	Razem
Nie	91	208	299
Kolumna %	85.05%	58.59%	
Tak	16	147	163
Kolumna %	14.95%	41.41%	
Razem	107	355	462

Wyniki przeprowadzonych testów niezależności chi-kwadrat pozwalają wnioskować, że stosunkowo najsilniejsze czynniki podwyższające ryzyko choroby wieńcowej to: występowanie choroby serca w rodzinie badanego, starszy wiek, podwyższone ciśnienie krwi oraz większa otyłość.

W drugiej części analizy dla oceny łącznego wpływu branych pod uwagę czynników jakościowych i ilościowych wykorzystano dwie różne techniki budowania modeli klasyfikacyjnych: **uogólnioną analizę dyskryminacyjną (GDA)** oraz **drzewa klasyfikacyjne**. Pierwsza z nich stanowi rozszerzenie klasycznej analizy dyskryminacyjnej, polegające na możliwości uwzględniania w charakterze predyktorów (zmiennych objaśniających) nie tylko zmiennych ilościowych, ale także zmiennych jakościowych oraz ich interakcji. Przy budowaniu modelu dla jakościowej zmiennej zależnej można stosować różne techniki włączania zmiennych.

Poniżej w tabeli zamieszczono wyniki uogólnionej analizy dyskryminacyjnej w postaci wielowymiarowych testów istotności, które pozwalają stwierdzić, które zmienne lub ich kombinacje zostały uwzględnione w modelu.



Efekt	Uogólniona analiza dyskryminacyjna Wielowymiarowe testy istotności					
	Test	Wartość statystyki	F	Efekt df	Błąd df	p
Wyraz wolny	Wilksa	0,9320	26,637	1	365	0,00000
Ciśn kat*Zach A kat*Pal kat	Wilksa	0,9805	7,267	1	365	0,00735
{13}Palenie	Wilksa	0,9807	7,177	1	365	0,00772
1*3*4*6	Wilksa	0,9812	6,984	1	365	0,00858
1*2*3*4*5*6	Wilksa	0,9819	6,737	1	365	0,00982
1*2*4*5*6*7	Wilksa	0,9841	5,902	1	365	0,01560
{7}Pal kat	Wilksa	0,9855	5,372	1	365	0,02102
Chor serca w rodz*Pal kat	Wilksa	0,9860	5,166	1	365	0,02361
{8}Wiek	Wilksa	0,9865	4,998	1	365	0,02598
Chol kat*Pal kat	Wilksa	0,9871	4,784	1	365	0,02936
Chor serca w rodz*Otył kat	Wilksa	0,9874	4,660	1	365	0,03152
Chor serca w rodz*Wiek kat*Pal kat	Wilksa	0,9876	4,586	1	365	0,03289
Wiek kat*Ciśn kat*Otył kat	Wilksa	0,9885	4,264	1	365	0,03963
1*4*5*7	Wilksa	0,9885	4,243	1	365	0,04012
Chor serca w rodz*Wiek kat	Wilksa	0,9886	4,206	1	365	0,04099
Ciśn kat*Otył kat*Zach A kat	Wilksa	0,9887	4,172	1	365	0,04183
2*3*4*5	Wilksa	0,9889	4,098	1	365	0,04366
Ciśn kat*Zach A kat	Wilksa	0,9894	3,913	1	365	0,04866
Chor serca w rodz*Ciśn kat*Chol kat	Wilksa	0,9895	3,891	1	365	0,04930

Otrzymany model jest dość rozbudowany i oprócz pojedynczych zmiennych jakościowych i ilościowych uwzględni także kombinacje zmiennych jakościowych. Szczególną uwagę zwraca wpływ palenia, gdyż do modelu weszła zarówno zmienna oznaczająca liczbę wypalanych papierosów, jak i jakościowa zmienna (Pal kat) zawierająca tylko informację, czy dana osoba paliła czy też nie. Oprócz tego, zmienna ta została uwzględniona w modelu w kombinacji z innymi zmiennymi jakościowymi, np. Ciśn kat, Zach A kat czy Chol kat. Oznacza to, że współwystępowania niektórych czynników jakościowych może podwyższać ryzyko wystąpienia choroby serca.

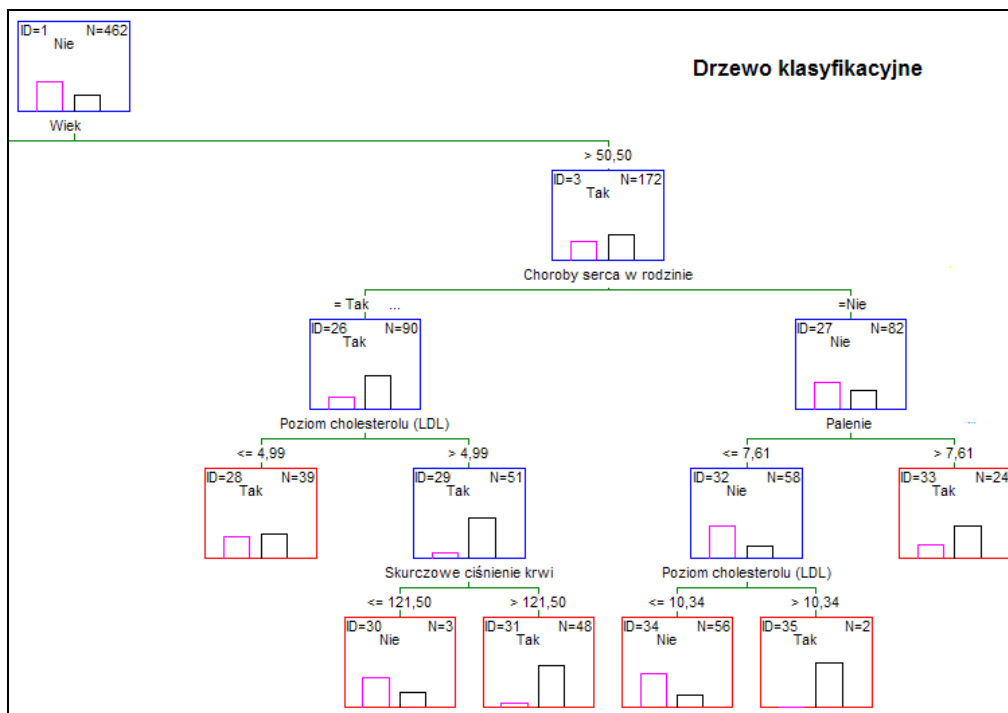
Jednym z najczęściej stosowanych kryteriów oceny jakości zbudowanego modelu jest odsetek poprawnych klasyfikacji. Dla otrzymanego modelu został on przedstawiony poniżej.

Klasa	Macierz klasyfikacji		
	Procent poprawnych	Nie p=,6472	Tak p=,3528
Nie	83,61%	250	49
Tak	69,34%	50	113
Razem	78,57%	300	162

Na podstawie tych wyników możemy stwierdzić, że model poprawnie klasyfikuje ponad 83% badanych, u których nie występowała choroba serca oraz blisko 70% tych, u których choroba serca rzeczywiście wystąpiła.

Drugą zastosowaną metodą budowania modelu były drzewa klasyfikacyjne (Gatnar 2001, Hastie i wsp. 2001). W metodzie tej cała przestrzeń zmiennych jest stopniowo dzielona na rozłączne podzbiory, aż do momentu uzyskania ich jednorodności ze względu na badaną zmienną zależną. Wyniki są najczęściej przedstawiane w postaci tzw. drzewa decyzyjnego, które umożliwia podanie reguł podziału na jednorodne grupy obiektów.

Poniżej przedstawiono fragment otrzymanego drzewa.



Na jego podstawie możemy podać przykładową regułę klasyfikacji:

Zawał serca wystąpił u 92% badanych mężczyzn powyżej 50,5 lat, u których występowały Choroby serca w rodzinie, poziom cholesterolu przekraczał 5 mmol/l, a ciśnienie krwi przekraczało 121,5 mm Hg.

Podobnie jak poprzednio, przy ocenie jakości modelu możemy posłużyć się odsetkiem poprawnych klasyfikacji. Został on przedstawiony w poniższej tabeli.

	Macierz klasyfikacji			Wiersze Razem
	Obserwowane	Przewidywane Nie	Przewidywane Tak	
Liczba	Nie	249	50	299
Kolumna %		84.69%	29.76%	
Liczba	Tak	45	118	163
Kolumna %		15.31%	70.24%	
Razem %		63.64%	36.36%	

Jak widać, otrzymany model daje wyniki podobne do tych, które uzyskano za pomocą uogólnionej analizy dyskryminacyjnej, i pozwala poprawnie sklasyfikować ponad 70 % badanych, u których wystąpił zawał serca, oraz blisko 85 % badanych, u których zawał nie wystąpił.

Wyniki analizy tego typu mogą zostać wykorzystane przy profilaktyce chorób serca oraz w badaniach przesiewowych.



## Podsumowanie

Dobra znajomość metod analizy danych stanowi niezbędne instrumentarium badacza rozwiązującego problemy w oparciu o dane empiryczne i powinna stanowić ważny element jego świadomości metodologicznej (Rao 1994, Tadeusiewicz 2000). Coraz bardziej powszechnemu wykorzystaniu różnych technik analizy danych sprzyja niezwykle intensywny rozwój nowych metod opartych na „brutalnej mocy” obliczeniowej współczesnych komputerów (Durka 2003). Jednocześnie metody te są bardzo łatwo dostępne w ramach specjalistycznego oprogramowania do analizy danych (dobrym przykładem jest tutaj program *STATISTICA*). Trzeba jednak na koniec jeszcze raz wyraźnie podkreślić, że nawet najbardziej wyszukane techniki analityczne nie zwalniają badacza od konieczności poprawnego zaplanowania badań, starannego doboru jednostek do badań, dokładnego sprawdzania założeń stosowanych metod, przestrzegania klasycznych schematów wnioskowania oraz ostrożnego formułowania wniosków z przeprowadzonych badań. Żadna bowiem metoda nie jest w stanie w sposób automatyczny przekształcić danych liczbowych w wiedzę naukową.

## Literatura

1. Cobb G., W., 1998, Introduction to Design and Analysis of Experiments, Springer.
2. Durka P. J., 2003, Wstęp do współczesnej statystyki, Wydawnictwo Adamantan.
3. Gatnar E., 2001, Nieparametryczna metoda dyskryminacji i regresji, PWN.
4. Giudici P., 2003, Applied Data Mining, Wiley.
5. Hastie T., Tibshirani R., Friedman J., 2001, The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer-Verlag, New York.
6. Moczko J., 2003, Wybrane metody eksploracji danych i wspomaganie procesów decyzyjnych, w: „Metody statystyki i data mining w badaniach naukowych”, StatSoft Polska.
7. Rao C., R., 1994, Statystyka i prawda, PWN, Warszawa.
8. Sokołowski A., 2002, Wprowadzenie do zastosowań metod statystycznych i technik data mining w badaniach naukowych, w: „Metody statystyczne i techniki data mining w badaniach naukowych”, StatSoft Polska.
9. Tadeusiewicz R., 2000, Drogi i bezdroża statystyki w badaniach naukowych, w: „Statystyka w badaniach naukowych”, StatSoft Polska.