Pakiet zautomatyzowanej analizy statystycznej PC SSTAT

1. Wprowadzenie

PC SSTAT jest oryginalnym i sprawdzonym oprogramowaniem do zautomatyzowanej analizy statystycznej na komputerze personalnym typu IBM PC.

Pakiet umożliwia rozwiązywanie różnorodnych problemów statystycznych, od statystyki opisowej (średnia, mediana, moda, odchylenie standardowe itp.), do wspomagania eksploracji danych (data mining) i tworzenia różnorodnych modeli matematycznych w oparciu o wyniki analiz wielowymiarowych.

Użytkownikami pakietu mogą być zarówno osoby zaawansowane w stosowaniu metod statystycznych, jak i te, które do tej pory nie stosowały takich metod z powodu braku przystępnego oprogramowania lub zbyt małej wiedzy z zakresu statystyki matematycznej i informatyki.

Pierwsze programy pakietu powstały dla komputera ODRA 1304 w roku 1980. Kolejne wersje pakietu były opracowywane od roku 1984 dla komputerów linii SM (PDP-11). Od roku 1986 pakiet jest rozwijany wyłącznie dla komputerów typu IBM PC, najpierw dla systemu operacyjnego DOS, a następnie dla systemu operacyjnego Windows.

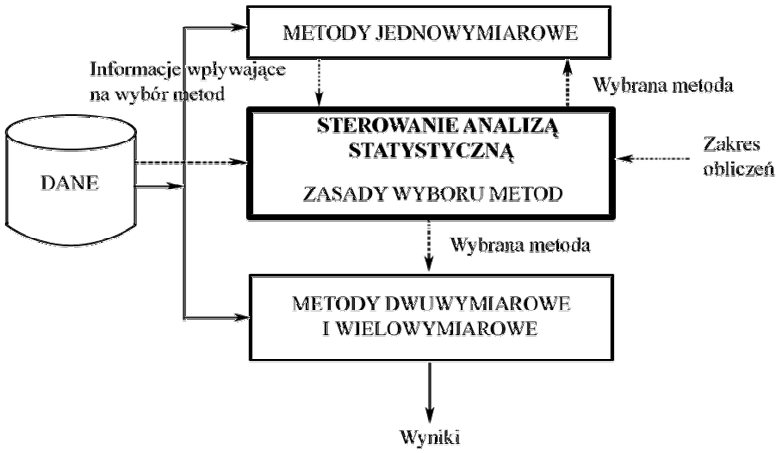
Koncepcję pakietu opracowano ze szczególnym uwzględnieniem konieczności przestrzegania warunków stosowania poszczególnych metod, bowiem tylko przy ich spełnieniu otrzymywane wyniki analizy statystycznej są poprawne. W tym miejscu należy podkreślić, że inne pakiety nie wprowadzają takich ograniczeń, umożliwiając wybranie do obliczeń oprogramowanych metod, np. testów Studenta i współczynnika korelacji Pearsona, bez sprawdzania obowiązujących je założeń.

2. Stosowane mechanizmy automatyzacji

W czasie dotychczasowej eksploatacji pakiet okazał się efektywnym i chętnie stosowanym narzędziem w pracy naukowo-badawczej. Stan taki nie uległ zmianie, mimo dostępności dla komputerów personalnych szeregu firmowych pakietów statystycznych. Spowodowane jest to m.in. wymienionymi poniżej rozwiązaniami, przyjętymi przy projektowaniu pakietu:

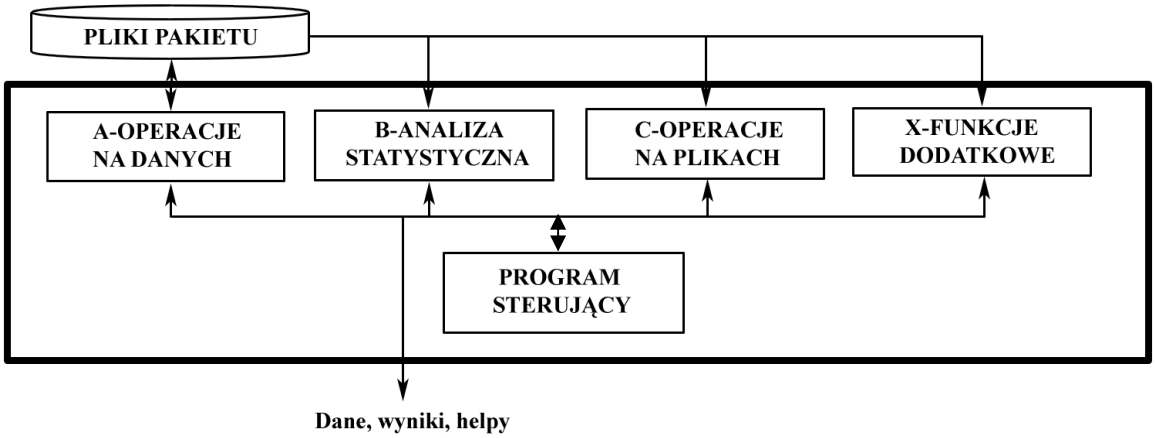
- Automatyzacją operacji na danych i określania zakresu analizy, zapewniającą radykalne ograniczenie pracochłonności i bardzo krótki czas wykonywania analiz;
- Automatycznym wyborem właściwych parametrów, współczynników i testów, zwłaszcza podczas analizy dwuwymiarowej, gwarantującym prawidłowość wykonywanych analiz;
- Automatycznym obliczaniem poziomów istotności podczas wykonywania wszystkich analiz, ułatwiającym dokonanie precyzyjnej interpretacji wyników.

Istotę automatyzacji w zakresie analizy statystycznej przedstawia poniższy rysunek.



Rysunek 1. Istota automatyzacji analizy statystycznej

W strukturze pakietu występuje program sterujący oraz cztery klasy funkcji:



Rysunek 2. Struktura pakietu SSTAT4

Poniżej krótko opisano każdą z powyżej wymienionych klas funkcji.

3. Operacje na danych

Operacje na danych umożliwiają między innymi: wprowadzanie danych przy pomocy klawiatury, import danych z pliku tekstowego, łączenie danych, wydzielanie podzbiorów danych, przekształcanie danych oraz eksport danych do pliku tekstowego.

Analiza statystyczna

4. Uwzględnione w pakiecie PC SSTAT metody statystyczne przedstawia tabela 1.

W pakiecie następuje automatyczny wybór metod w ramach poniżej podanych analiz:

- Analiza jednowymiarowa - Tab. 2.
- Ocena istotności zależności statystycznej i korelacji dwóch cech - Tab. 3.
- Ocena istotności różnic rozkładów cechy przy dwóch warunkach - Rys. 3.
- Ocena istotności różnic rozkładów cechy przy wielu warunkach - Tab. 4.

Tabela 1. Wskaźniki i metody statystyczne

<table>
<tr><th rowspan="2">Liczba grup danych</th><th colspan="3">Liczba cech</th></tr>
<tr><th>1</th><th>2</th><th>≥ 2</th></tr>
<tr><td>1</td><td>ANALIZA JEDNOWYMIAROWA
Błędy grube
Centyle
Estymacja parametrów rozkładu
Ocena normalności
Ocena losowości</td><td>ANALIZA DWYWYMIAROWA
Test niezależności
Współczynnik Pearsona
Współczynnik Spearmana
Współczynnik Cramera</td><td>ANALIZA WIELOWYMIAROWA
Regresja liniowa, wielomianowa i potęgowa
Analiza czynnikowa
Analiza skupień dla cech i obiektów
Korelacja kanoniczna</td></tr>
<tr><td>2</td><td>ANALIZA DWUWYMIAROWA
Testy Studenta
Test Cochrana-Coxa
Test rangowanych znaków
Test Wilcoxona
Test dokładny Fishera
Test McNemara
Test chi-kwadrat</td><td colspan="2">ANALIZA WIELOWYMIAROWA
Wielowymiarowa analiza wariancji i analiza dyskryminacji</td></tr>
</table>

Tabela 2. Zakres analizy jednowymiarowej w zależności od rodzaju skali

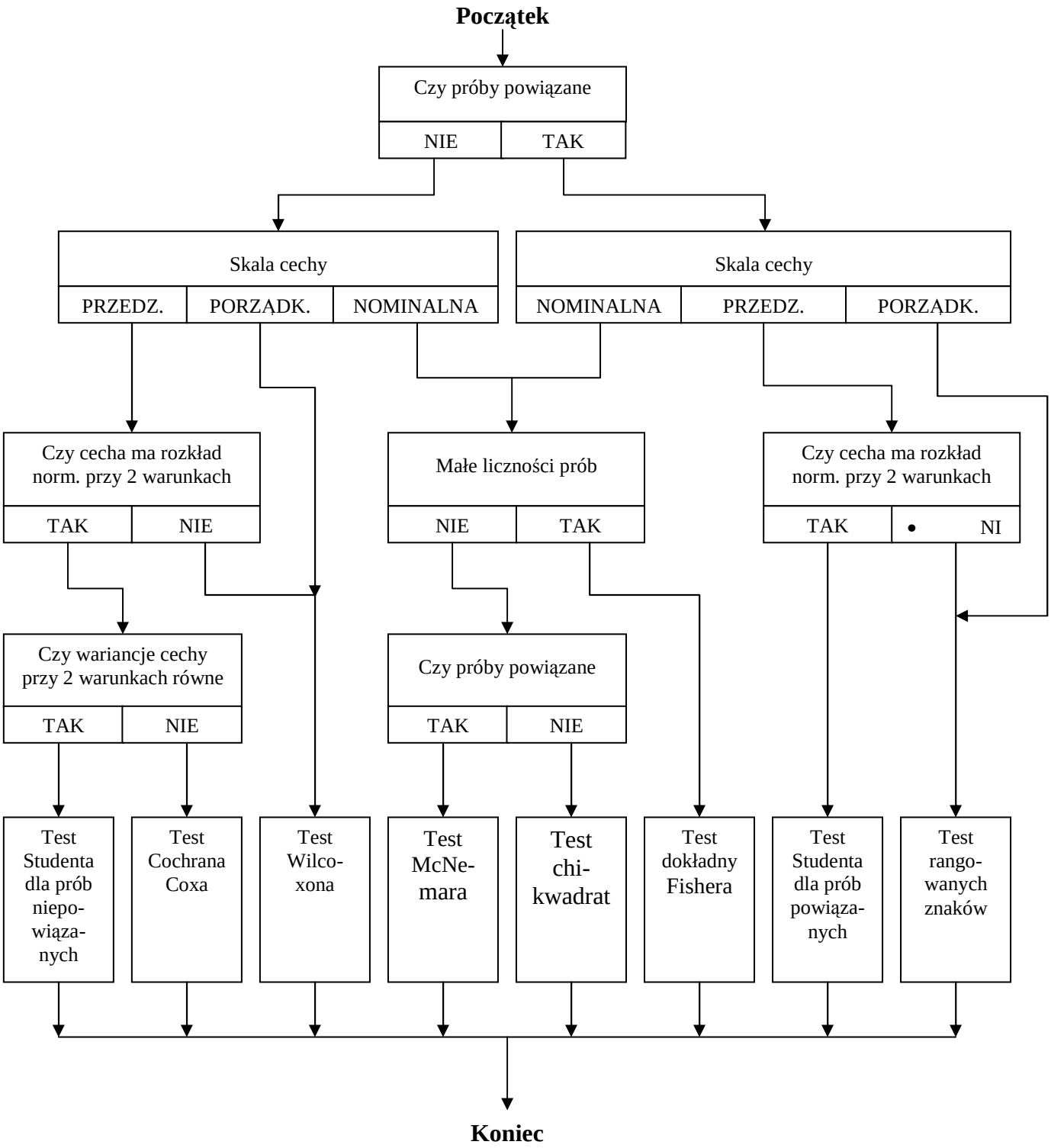
Lp.	Wskaźniki i testy analizy jednowymiarowej	Skala		
		przedziałowa	porządkowa	nominalna
1.	Estymacja punktowa i przedziałowa parametrów rozkładu	X	X	X
2.	Weryfikacja hipotez dotyczących parametrów rozkładu	X	X	X
3.	Ocena losowości pobrania próby	X	X	X
4.	Ocena normalności rozkładu	X		
5.	Wyznaczanie wentyli	X		
6.	Eliminacja błędów grubych	X		
7.	Zastępowanie elementów brakujących średnią	X		
8.	Zastępowanie elementów brakujących medianą	X	X	
9.	Zastępowanie elementów brakujących modą			X

Tabela 3. Zasady wyboru współczynników w zależności od skali cech

<table>
<tr><th rowspan="2">I cecha</th><th colspan="4">II cecha</th></tr>
<tr><th>Skala przedziałowa rozkład normalny</th><th>Skala przedziałowa rozkład różny od normalnego</th><th>Skala porządkowa</th><th>Skala nominalna</th></tr>
<tr><td>Skala przedziałowa rozkład normalny</td><td>PEARSON</td><td colspan="2" rowspan="3">SPEARMAN</td><td rowspan="3">CRAMER</td></tr>
<tr><td>Skala przedziałowa rozkład różny od normalnego</td></tr>
<tr><td>Skala porządkowa</td></tr>
<tr><td>Skala nominalna</td><td colspan="3"></td></tr>
</table>

W pakiecie zapewniono automatyczne eliminowanie elementów brakujących w trakcie analizy jednowymiarowej i dwuwymiarowej oraz kontrolę występowania takich elementów na początku każdej analizy wielowymiarowej, przy czym elementy brakujące można zastąpić wskazanym wskaźnikiem, np. średnią lub medianą albo wybrać do analizy obserwacje bez elementów brakujących. Drukowanie lub zapamiętywanie generowanych na ekranie monitora kolorowych wykresów możliwe jest przy wykorzystaniu standardowych programów typu „screen capture”.

Na poniższym rysunku przedstawiono schemat blokowy wyboru testów do oceny istotności różnic rozkładu określonej cechy w dwóch warunkach, spośród opisanych w niniejszym punkcie. Wszystkie te metody zostały opisane w dotychczasowych rozważaniach.



Rysunek 3. Schemat blokowy wyboru testów statystycznych do oceny istotności różnic rozkładu cechy w dwóch różnych warunkach

Tabela 4. Zasady wyboru testów do oceny istotności rozkładu cechy

Nazwa testu	Próby powiązane		Skala cechy			Rozkład normalny cechy w k populacjach		Równość wariancji cechy w k populacjach	
	T	N	Prz.	Porz.	Nom.	T	N	T	N
Analiza wariancji		X	X			X		X	
Test $q_{\overline{X}}$	X		X			X			X
Kruskala-Wallisa		X	X			X			X
		X	X				X		
		X		X					
Friedmana	X		X				X		
	X			X					
Góralskiego		X			X				
Cochrana	X				X				

Charakterystyka metod wielowymiarowych oprogramowanych w pakiecie jest następująca:

Analiza regresji wykorzystywana jest do szukania związku funkcyjnego pomiędzy tzw. zmienną zależną i określoną liczbą tzw. zmiennych niezależnych. Najczęściej przyjmuje się związek liniowy. W przypadku małej liczby zmiennych niezależnych, szuka się też związku w postaci wielomianu. Możliwe jest ustalenie a priori zmiennych niezależnych, które ujmowane są w równaniu regresji lub też określenie tylko ich zbioru. W tym przypadku, do równania wprowadzane są tylko te zmienne, które charakteryzuje określony współczynnik korelacji cząstkowej ze zmienną zależną.

Analiza czynnikowa pozwala na podział analizowanych zmiennych na określoną liczbę grup, z których każda kształtowana jest samoistnie przez oddzielny czynnik.

Analiza korelacji kanonicznej wykorzystywana jest do wyznaczania związku liniowego pomiędzy dwiema grupami zmiennych. Można traktować ją więc jako uogólnienie analizy regresji.

Analiza skupień wykorzystywana jest do podziału zbioru określonych elementów na grupy, których obiekty są „podobne" do siebie. Obiektami mogą być zarówno dowolne elementy materialne opisane wybranymi cechami, jak i cechy opisujące rozpatrywane elementy materialne.

Wielowymiarowa analiza wariancji (MANOVA) wykorzystywana jest do weryfikacji hipotez o równości kilku wektorów wartości oczekiwanych. Jest ona rozszerzeniem analizy wariancji (ANOVA), albowiem rozpatruje powyższą hipotezę dla kilku wartości oczekiwanych. MANOVA stosowana jest w powiązaniu z **analizą dyskryminacji**, której ważnym krokiem jest zastąpienie wielu cech naturalnych małą liczbą zmiennych abstrakcyjnych bez zmniejszenia zróżnicowania grup. Możliwe jest też wybranie cech najbardziej różnicujących. W ramach tej analizy prowadzona jest klasyfikacja na podstawie cech abstrakcyjnych. Stopień jej zgodności z podziałem a priori świadczy poglądowo o występującym zróżnicowaniu grup.

5. Operacje na plikach i funkcje dodatkowe

Operacje na plikach umożliwiają: edycję, kasowanie, kopiowanie i drukowanie zawartości tworzonych plików.

Funkcje dodatkowe obejmują testowe sprawdzania wiadomości z podstaw statystyki oraz monitorowanie wykorzystywania pakietu.

6. Podsumowanie

Pakiet PC SSTAT został skonstruowany w oparciu o ponad trzydziestoletnie doświadczenia dydaktyczne i naukowo-badawcze, zgromadzone podczas wykładów w kilku wyższych uczelniach oraz wykonywania analiz statystycznych wyników różnorodnych badań i eksperymentów. Pakiet ma także za sobą ponad dwudziestoletnią eksploatację użytkową w rozwiązywaniu najrozmaitszych problemów. W pakiecie zastosowano wcześniej opisane mechanizmy automatyzacji. Korzystanie z pakietu nie wymaga szczegółowej wiedzy ze statystyki i informatyki.

Podane powyżej cechy pozwalają zarekomendować pakiet PC SSTAT jako sprawdzone, łatwe w użyciu i efektywne narzędzie w pracy naukowo-badawczej i dydaktycznej, umożliwiające rozwiązywanie większości problemów we wszystkich możliwych obszarach tych działalności.