



## STATYSTYKA W MEDYCYNIE

*Arkadiusz Miziołek, StatSoft Polska Sp. z o.o.*

### Wprowadzenie

Nowoczesna medycyna nie może obejść się bez statystyki, co więcej, zakres zastosowań analizy danych w medycynie stale się poszerza. Techniki statystyczne pozwalają uzyskać odpowiedzi na wiele istotnych pytań, ustalać fakty, ujawniać prawidłowości i zależności, których w inny sposób w ogóle nie byłibyśmy w stanie wykryć.

Analiza danych jest wykorzystywana już nie tylko przez naukowców i lekarzy bezpośrednio zaangażowanych w badania kliniczne, lecz także tych lekarzy, którzy śledzą wyniki nowych badań i chcą poprawnie interpretować ich wyniki. W coraz większej liczbie publikacji przedstawiających wyniki badań medycznych stosowane są oprócz klasycznych metod statystycznych, takich jak analiza przeżycia czy testy statystyczne, również zaawansowane techniki data mining, takie jak sztuczne sieci neuronowe i drzewa decyzyjne.

### Obszary zastosowań analizy danych w medycynie

Techniki statystyczne niosą ze sobą bardzo wiele korzyści nie tylko w medycznych obszarach badawczych, ale również w administrowaniu jednostkami służby zdrowia, wpływają na polepszenie jakości opieki nad chorym i efektywność wykorzystania posiadanych zasobów.

Statystyka i data mining to dziedziny nauki, które przy wykorzystaniu odpowiednich narzędzi informatycznych uwalniających od żmudnego, ręcznego przetwarzania danych, stają się potężnym instrumentem do lepszego poznawania badanych zjawisk.

**Projektowanie badań.** Zaprojektowanie badania jest niezwykle ważnym etapem, ponieważ nawet najbardziej zaawansowane metody statystyczne nie są w stanie wyeliminować błędów popełnionych na etapie planowania badań. W tym przypadku narzędzia dostępne w programach *STATISTICA* mogą pomóc w doborze grupy kontrolnej czy liczności próby.

**Wspomaganie decyzji diagnostycznych.** Ten obszar zastosowań analizy danych ma wciąż ogromny potencjał rozwojowy. W tym przypadku metody statystyczne i data mining mogą być niezwykle przydatne, m.in. do szacowania dokładności testów diagnostycznych, porównania ze sobą konkurencyjnych testów, jak również w odkrywaniu nowych reguł



diagnostycznych. Przedmiotem analiz mogą być zarówno pojedyncze czynniki rokownicze – parametry opisujące stan pacjenta, jak i złożone, wielowymiarowe dane medyczne (np. zapis EKG, zdjęcia USG, dane laboratoryjne) czy też nawet dane genetyczne.

**Stosowanie nowej terapii/leku.** Zmiana procedur leczenia, stosowanie nowej terapii czy leku muszą być w odpowiedni sposób uzasadnione badaniami empirycznymi. Techniki statystyczne, a coraz częściej również metody data mining pełnią bardzo istotną funkcję w poszukiwaniu nowych terapii czy leków. Pozwalają bowiem nie tylko oddzielić to co przypadkowe, od tego co przyczynowe, ale również prowadzą do ustalenia nowych zależności, niemożliwych do ustalenia bez wykorzystania tych metod.

**Podejmowanie decyzji w zakresie prewencji i chorób.** Wykorzystując narzędzia analityczne dostępne w *STATISTICA* możemy z dużym prawdopodobieństwem określać ryzyko wystąpienia u poszczególnych pacjentów danego zjawiska (np. zawału lub udaru mózgu) lub wyodrębnić grupy pacjentów obciążonych podwyższonym ryzykiem i dzięki temu poprawić trafność wyborów strategii leczenia i profilaktyki.

**Analiza wyników badań klinicznych.** We współczesnej medycynie podstawowym narzędziem oceny skuteczności leków, procedur medycznych, testów diagnostycznych, modyfikacji czynników ryzyka, diety czy też efektów edukacji zdrowotnej są badania kliniczne. W zależności od celów badań i rodzaju stosowanych interwencji są one projektowane w standardzie badań eksperymentalnych lub obserwacyjnych. W badaniach klinicznych wyniki nigdy nie są jednoznaczne, statystyka pozwala wykazać, że otrzymane wyniki nie są dziełem przypadku. Program *STATISTICA* zawiera kompletny zestaw narzędzi i metod statystycznych służących do wszechstronnego opracowania i graficznej prezentacji wyników badań klinicznych.

**Badania genetyczne.** Duże ilości danych dostarczanych przez badania genetyczne wymagają niejednokrotnie specjalnego wieloetapowego podejścia. Niezbędny jest etap przygotowania danych do analizy, konieczne staje się stosowanie technik redukcji wymiaru, grupowania obiektów podobnych oraz selekcji istotnych cech. Wynikiem takiej wstępnej analizy są na przykład listy genów o istotnie zmienionej ekspresji, liczące i tak setki, a nawet tysiące elementów. W celu wyszukiwania wzorców w tak dużych bazach danych często wykorzystywane są specjalne algorytmy data mining nastawione na analizę dużych ilości wielowymiarowych danych i uwzględnianie bardzo złożonych zależności.

**Metaanaliza.** Ilościowa synteza wyników pochodzących z niezależnych badań pierwotnych pozwala rozszerzyć wnioski z pojedynczych badań na szerszą populację i bardziej precyzyjnie ocenić wpływ działania analizowanego czynnika. Procedury statystyczne dostępne w programie *STATISTICA Zestaw Medyczny* umożliwia m.in. wykonywanie obliczeń dla szeregu miar efektu, przedstawienie wyników metaanalizy w postaci szczegółowego raportu oraz wykresu typu *forest plot*. Narzędzie pozwala ocenić niejednorodność badań, przeprowadzić analizę w grupach i analizę wrażliwości efektu łącznego na zmiany modelu.

**Zarządzanie.** Odpowiednio zaprojektowane systemy analityczne wdrażane w skali całej kliniki lub szpitala nie tylko są wykorzystywane do wspomaganie podejmowania decyzji medycznych, prognozowania wykonania kontraktów, ale również informują o efektyw-



ności wykorzystania posiadanych zasobów (finansowych, materialnych i personalnych) dzięki czemu przyczyniają się do poprawy rachunku ekonomicznego.

**Automatyzacja sprawozdawczości.** Narzędzia analizy danych pomagają zarówno w automatyzacji raportowania rutynowych działań, jak również w monitorowaniu stanu realizacji projektów i badań medycznych, uwalniając tym samym lekarzy od wielu czasochłonnych zajęć.

**Text mining.** W oparciu o techniki statystyczne możemy uwzględniać w analizie informacje zapisane w języku naturalnym np. rozpoznania, wywiady, opisy wyników badań medycznych itp. Możliwe jest zatem powiązanie nieustrukturyzowanego tekstu i danych sformalizowanych w celu pozyskiwania nowych, interesujących nas informacji.

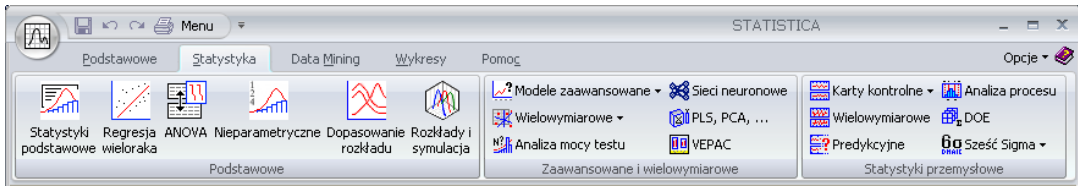
## Rozwiązania StatSoft

*STATISTICA* to rodzina programów zapewniająca kompletne rozwiązania w zakresie analizy danych – najlepsze narzędzia analityczne, systemy korporacyjne z elementami pracy grupowej i kontroli dostępu oraz możliwością zdalnej pracy przez Internet, mechanizmy dostępu do źródeł danych oraz systemy rozpowszechniania informacji. Wybierając poszczególne produkty z rodziny *STATISTICA*, można zbudować rozwiązanie dostosowane do indywidualnych potrzeb użytkowników.

*STATISTICA* spełnia wszystkie wymagania, jakie osoby wykorzystujące statystykę w naukach biomedycznych stawiają systemom analitycznym:

- ◆ dostarcza proste w obsłudze, intuicyjne środowisko pracy;
- ◆ zapewnia otwartość i elastyczność umożliwiającą wymianę danych pomiędzy różnymi aplikacjami, w tym MS Excel;
- ◆ udostępnia komplet zarówno prostych, jak i zaawansowanych narzędzi analizy danych wykorzystywanych do opracowywania wyników badań medycznych – klasyczne metody statystyczne, jak i specjalistyczne techniki data mining;
- ◆ zawiera specjalnie zaprojektowane procedury statystyczne powszechnie wykorzystywane w naukach medycznych;
- ◆ ma ogromne możliwości wizualizacji danych, umożliwiające tworzenie atrakcyjnych graficznie raportów i zestawień.

Zakres metod analitycznych oferowanych przez *STATISTICA* jest niezwykle szeroki, można z nich zbudować odpowiedni zestaw analityczny dopasowany do indywidualnych lub specyficznych potrzeb użytkownika. Poszczególne programy z rodziny *STATISTICA* można łączyć w jednym pakiecie, który może z czasem być rozszerzany o dodatkowe moduły. Dzięki temu możemy mieć dostęp do bardzo wielu różnych metod analitycznych zawartych w jednej aplikacji *STATISTICA*.

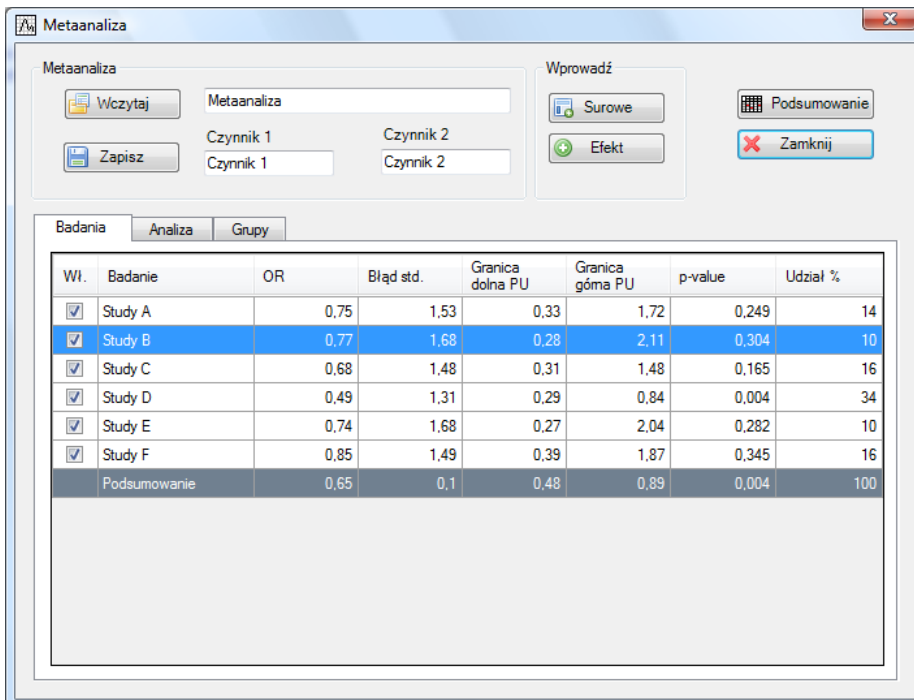


Poniżej znajdują się opisy wybranych aplikacji *STATISTICA*, najczęściej wykorzystywanych do opracowywania wyników badań medycznych.

### ***STATISTICA Zestaw Medyczny***

Wychodząc naprzeciw oczekiwaniom osób zajmujących się opracowywaniem wyników badań medycznych, opracowaliśmy specjalistyczny pakiet analityczny *STATISTICA Zestaw Medyczny*, który stanowi uzupełnienie funkcjonalności programów z rodziny *STATISTICA*. W skład *Zestawu Medycznego* wchodzi: *Moduł Metaanaliza*, *Miary powiązania/efektów*, *Krzywe ROC*, *Wykres Blanda-Altmana* oraz *Test post hoc ANOVA Friedmana*.

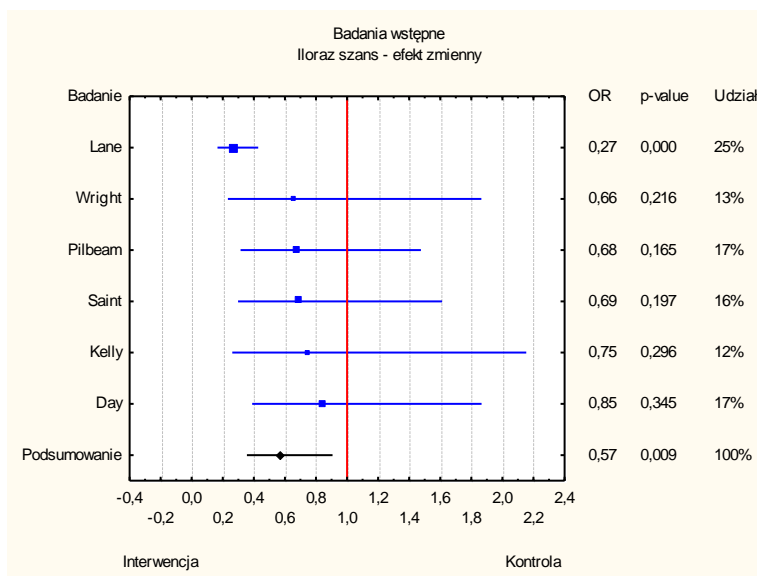
*Metaanaliza* jest narzędziem umożliwiającym syntezę wyników pochodzących z wielu badań pierwotnych i obliczanie na ich podstawie łącznej miary efektu. Pozwala bardziej precyzyjnie ocenić wpływ działania analizowanego czynnika oraz rozszerzyć wnioski z pojedynczych badań na szerszą populację. Obliczenia są wykonywane dla szeregu miar efektu, a wyniki można przedstawić w postaci szczegółowego raportu oraz wykresu typu *forest plot*.



W analizie można uwzględnić zarówno sytuację, w której różnica efektów dla rozważanych badań jest związana jedynie z błędem losowym (*common effect model*), jak i model, w którym na różnicę efektów wpływają dodatkowo inne charakterystyki poszczególnych badań (*random effect model*). Dzięki temu możemy bardziej precyzyjnie ocenić wpływ działania analizowanego czynnika oraz rozszerzyć wnioski z pojedynczych badań na szerszą populację.

Moduł *Metaanaliza* umożliwia m.in.:

- ♦ wyliczenie miar efektu (np. różnica średnich, iloraz średnich, g Hedgesa, d Cohena, iloraz szans, ryzyko względne, różnica ryzyka, współczynnik korelacji) dla pojedynczych badań;
- ♦ uwzględnienie w analizie badań z wynikami przedstawionymi w postaci różnych miar efektu (np. ilorazu szans oraz d Cohena);
- ♦ obliczanie łącznych miar efektu dla modelu z efektem wspólnym i zmiennym (*common/random effect model*);
- ♦ utworzenie wykresu typu *forest plot*;
- ♦ wykonanie analizy jednorodności badań (miary Q,  $T^2$ ,  $I^2$ );
- ♦ przeprowadzenie analizy wrażliwości – analiza po wyłączeniu badania lub grupy badań;
- ♦ przeprowadzenie metaanalizy w grupach (*subgroup analyses*).



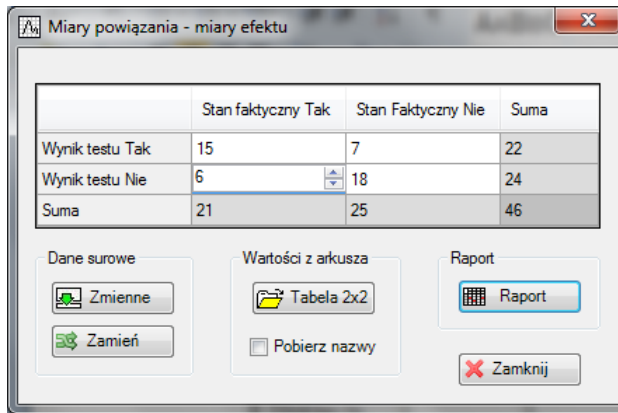
Interfejs użytkownika został zaprojektowany tak, aby w maksymalnym stopniu ułatwić badaczowi wykonywanie analiz poprzez:

- ♦ wprowadzenie wyników badań pierwotnych oraz już wyliczonych miar efektu ręcznie lub zapisanych w pliku *STATISTICA*;



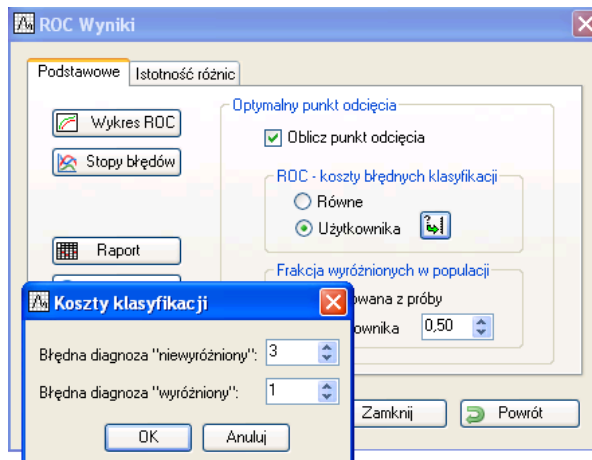
- ◆ automatyczny wybór badań adekwatnych dla wskazanej miary efektu;
- ◆ wygodny wybór pożądanego zakresu wyników;
- ◆ zapis projektu metaanalizy do pliku umożliwiającego kontynuację analiz lub uzupełnienie ich o dodatkowe badania w dowolnym czasie.

**Miary powiązania/efektów** jest dedykowanym modulem przeznaczonym do obliczania na podstawie tabeli 2x2 szeregu wskaźników powiązania lub efektu. Na przykład umożliwia on obliczenie efektu związanego z binarną zmienną zależną, spowodowanego manipulacją zmienną niezależną. Moduł umożliwia zarówno obliczenie wskaźników na podstawie danych surowych jak również ręczne wprowadzenie lub korektę wartości w tabeli.



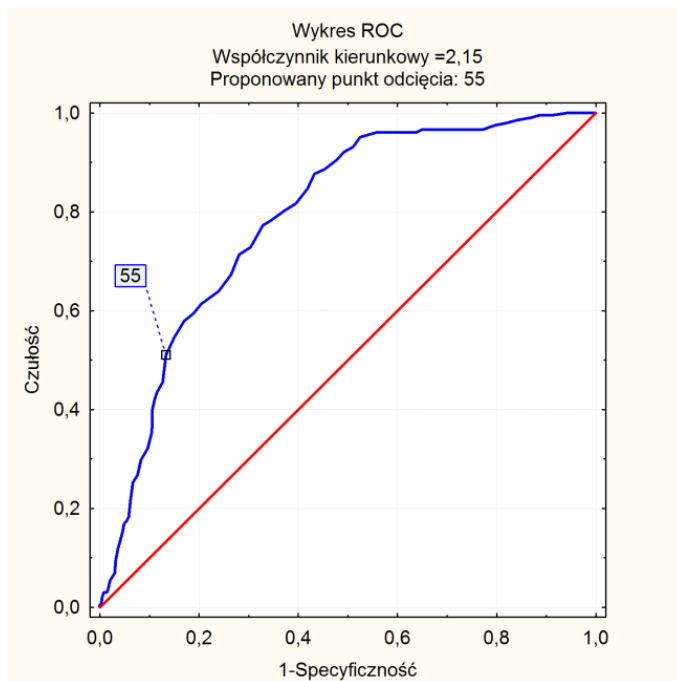
Miary te mają szczególne znaczenie w diagnostycznej analizie związków przyczynowych na podstawie rozkładów w tabeli obrazującej relacje typu test-efekt.

**Krzywe ROC (Receiver Operating Characteristic)** są narzędziem służącym do oceny poprawności klasyfikatora (pojedynczej zmiennej lub całego modelu), zapewniają one łączny opis jego czułości i specyficzności. Ten sposób wspomagania systemu decyzyjnego jest szeroko stosowany w różnych obszarach analizy danych m.in. w diagnostyce medycznej.



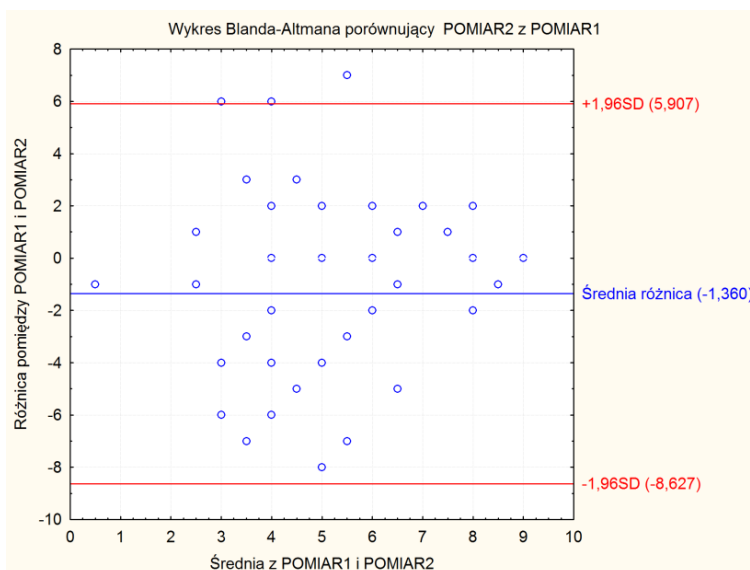
Moduł *Krzywe ROC* umożliwia:

- ◆ kreślenie krzywych ROC dla prób zależnych i niezależnych,
- ◆ obliczanie pola powierzchni pod krzywą,
- ◆ porównywanie istotności różnicy pól pomiędzy dwiema krzywymi
- ◆ ustalanie optymalnego punktu odcięcia dla podanych kosztów błędnej klasyfikacji i prawdopodobieństw a priori występowania badanego zjawiska,
- ◆ obliczanie miar FP, TP, FN, FP, *Sensitivity*, *Specificity*, ACC, PPV, NPV, *False positive ratio*, *False negative ratio*, LR dla wszystkich możliwych punktów odcięcia.



**Wykres Blanda-Altmana** wykorzystywany najczęściej w chemometrii i biostatystyce przedstawia stopień zgodności pomiędzy dwiema różnymi próbami bądź wskaźnikami. Merytorycznie jest on identyczny z wykresem średnia-różnica Tukeya.

Jednym z głównych zastosowań wykresu Blanda-Altmana jest porównanie dwóch wskaźników klinicznych, z których każdy obciążony jest pewnym błędem pomiaru. Może być on także wykorzystany do porównania nowej techniki pomiaru bądź wskaźnika z obowiązującym złotym standardem.



Na wykresie przedstawia się przedziały zgodności liczone jako średnia różnica pomiędzy badanymi wskaźnikami  $\pm 1,96 \cdot$  odchylenie standardowe różnicy.

**Kreator regresji logistycznej** umożliwia szybkie i wygodne wykonanie krok po kroku kompletnej analizy regresji logistycznej za pomocą intuicyjnego, przyjaznego dla użytkownika środowiska. Poza standardowymi, powszechnie stosowanymi opcjami dostępnymi dotychczas w programie *STATISTICA* kreator umożliwia:

- ♦ przygotowanie danych dla regresji – badanie liniowości wpływu zmiennych na zjawisko, analizę braków danych i współliniowości zmiennych, identyfikację interakcji,
- ♦ wykonanie jednoczynnikowej analizy regresji dla wszystkich zmiennych wraz z wykresem leśnym (*forest plot*),
- ♦ wykonanie regresji krokowej,
- ♦ zaawansowaną walidację modelu – krzywa ROC oraz walidacja za pomocą testu krzyżowego oraz metodą *leave one out* przydatną dla małych prób danych,
- ♦ zastosowanie techniki *propensity score*,
- ♦ zamianę zmiennych jakościowych na sztuczne zmienne zero-jedynkowe według wybranego schematu kodowania, co umożliwia uwzględnienie w modelowaniu zmiennych jakościowych.

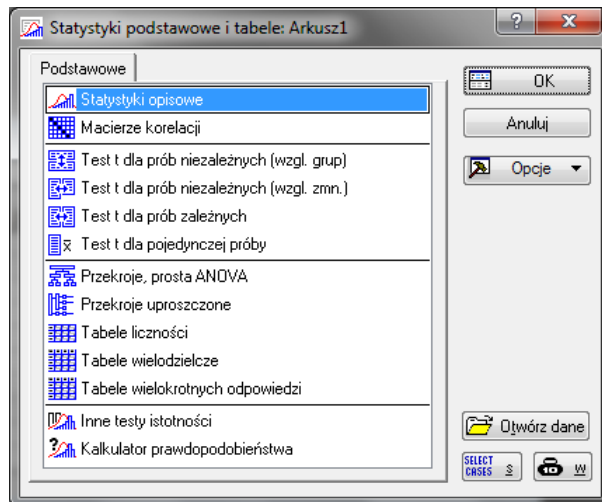


### **STATISTICA Pakiet Podstawowy PL**

*STATISTICA Pakiet Podstawowy* to obszerny zestaw najpopularniejszych metod statystycznych, obejmujący: *Statystyki opisowe*, *Tabele liczebności*, *Przekroje*, *Tabele wielodzzielcze*, *Korelacje*, *Testy parametryczne*, *ANOVA*, *Analiza regresji*, *Dopasowanie rozkładów*, *Testy nieparametryczne* i *Kalkulator prawdopodobieństwa*. Zawiera pełne środowisko




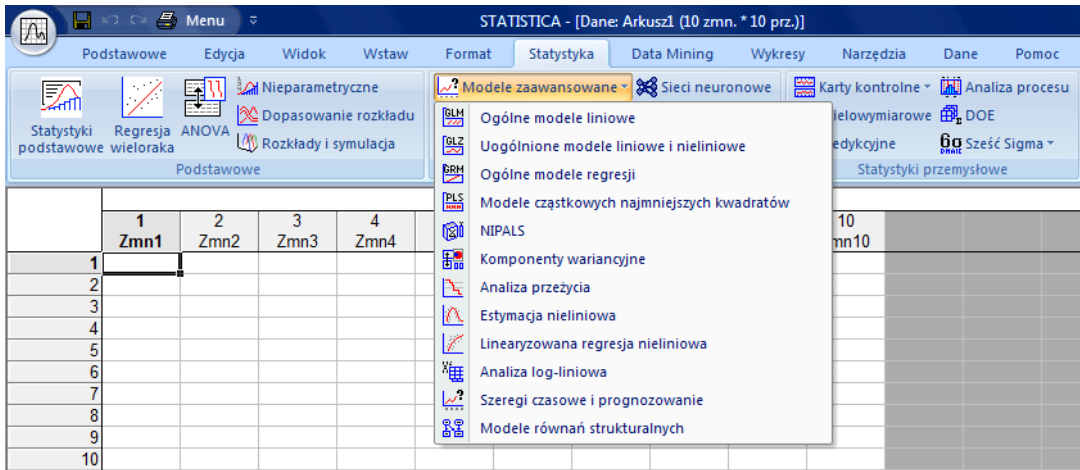
*STATISTICA*: modyfikowalny interfejs, elastyczne narzędzia zarządzania wynikami i raportowania, interakcyjne narzędzia do tworzenia zapytań do baz danych (*STATISTICA Query*), wszechstronne narzędzia importu i eksportu danych, komplet narzędzi wizualizacji danych i wyników analiz (wysokiej jakości wykresów) oraz w pełni zintegrowany język programowania *STATISTICA Visual Basic*.



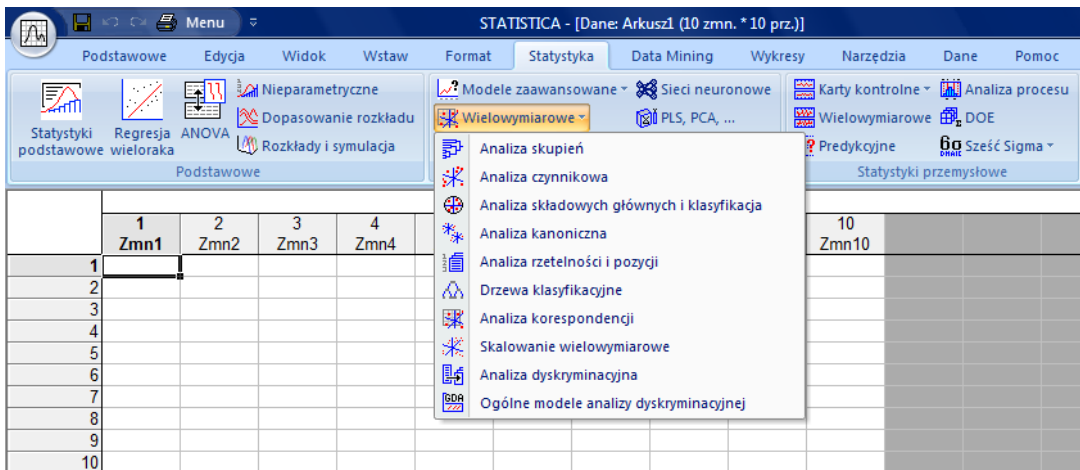
### ***STATISTICA Pakiet Zaawansowany PL***

Zestaw *STATISTICA Pakiet Zaawansowany* zawiera wszystkie funkcjonalności programu *STATISTICA Pakiet Podstawowy*, a także moduły *Modele Zaawansowane*, *Analizy Wielowymiarowe* oraz *Analiza Mocy Testów*.

 ***Modele Zaawansowane*** to szeroki wybór zaawansowanych narzędzi do budowy modeli liniowych i nieliniowych. Umożliwiają one wykorzystanie predyktorów ciągłych i jakościowych, uwzględnianie interakcji, tworzenie modeli hierarchicznych oraz automatyczny wybór modeli. Ponadto pozwalają obliczać komponenty wariancyjne, analizować dane o postaci szeregów czasowych oraz stosować wiele innych metod. Moduł obejmuje następujące grupy procedur: *Ogólne modele liniowe (GLM)*, *Uogólnione modele liniowe i nieliniowe*, *Ogólne modele regresji*, *Szeregi czasowe i prognozowanie*, *Modele częściowych najmniejszych kwadratów*, *Algorytm NIPALS (PCA/PLS)*, *Komponenty wariancyjne i model mieszany ANOVA/ANCOVA*, *Analiza czasu przeżycia i awaryjności*, *Estymacja nieliniowa*, *Linearyzowana regresja liniowa*, *Analiza log-liniowa tabel licznosci*, *Modelowanie równań strukturalnych i analiza ścieżkowa (SEPATH)*, *Rozkłady i symulacje*.



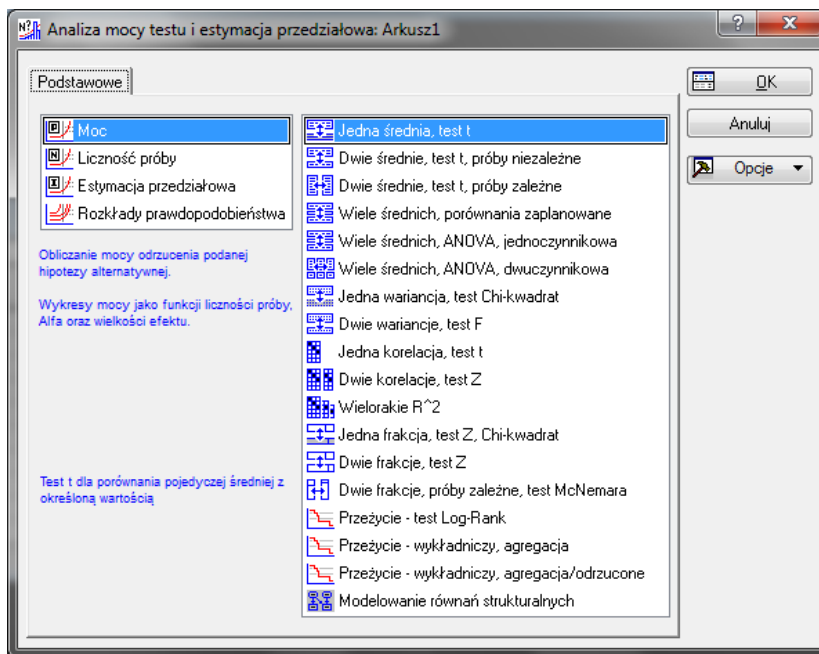
**Analizy Wielowymiarowe** to bogaty zestaw technik eksploracji danych różnego typu, począwszy od analizy skupień, a skończywszy na wybranych metodach drzew klasyfikacyjnych, wraz z obszernym zestawem interakcyjnych narzędzi wizualizacji, służących do badania współzależności i występowania regularnych układów danych. Moduł obejmuje następujące grupy procedur: *Analiza skupień*, *Analiza czynnikowa*, *Analiza składowych głównych i klasyfikacja*, *Analiza kanoniczna*, *Analiza rzetelności i pozycji*, *Drzewa klasyfikacyjne*, *Analiza korespondencji*, *Skalowanie wielowymiarowe*, *Analiza dyskryminacyjna*, *Ogólne modele analizy dyskryminacyjnej (GDA)*.



**Analiza Mocy Testów** to specjalistyczne, a jednocześnie przyjazne dla użytkownika narzędzie do analizowania wszelkich aspektów mocy testów statystycznych i wyznaczania liczności próby. Moduł ten obejmuje obszerny zestaw narzędzi ułatwiających takie zaplanowanie badań, aby wielkość próby odpowiadała ich celowi, a także analizy związane z obliczaniem mocy statystycznej i wielkości próby. *STATISTICA Analiza Mocy Testów*



wychodzi poza standardowe testy „braku różnic” i wykorzystuje także techniki bazujące na estymacji przedziałowej.



### **STATISTICA Automatyczne Sieci Neuronowe PL**

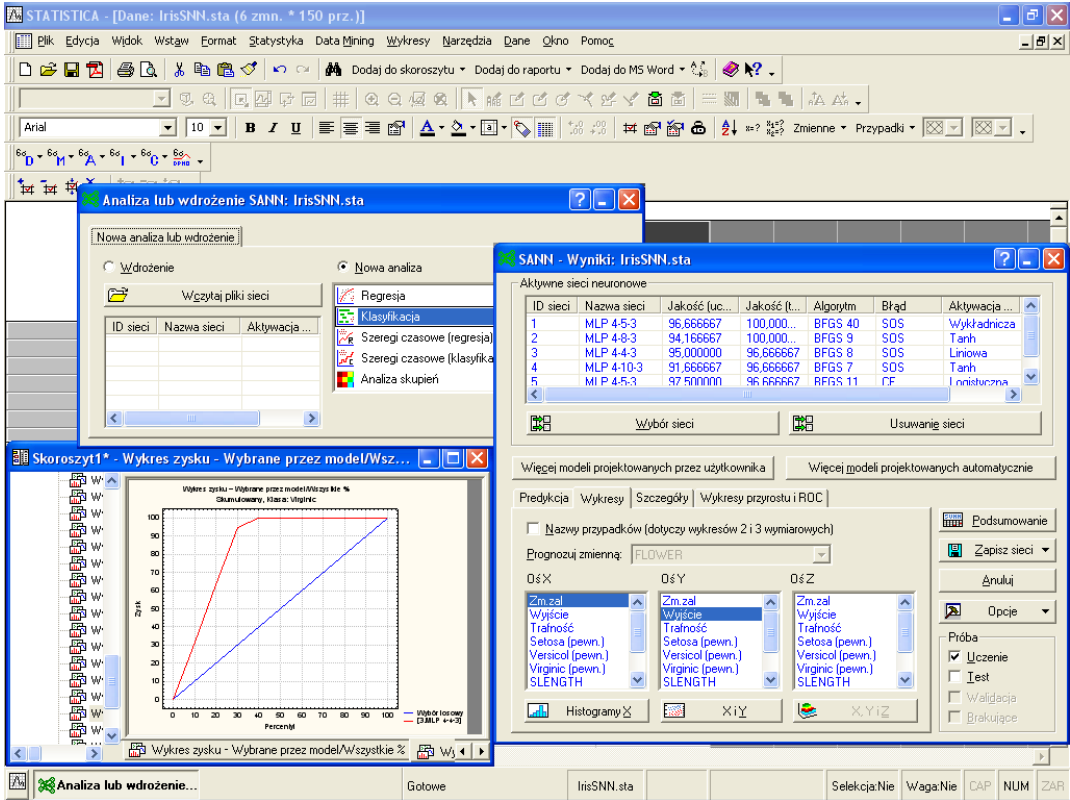
Program *STATISTICA Automatyczne Sieci Neuronowe* jest wydajną, a zarazem łatwą w obsłudze aplikacją wykorzystującą najnowszą technologię i algorytmy tworzenia oraz wdrażania modeli bazujących na sieciach neuronowych. Dzięki niej nawet użytkownicy niebędący specjalistami w zakresie sieci neuronowych mogą wykorzystywać te techniki sztucznej inteligencji do rozwiązywania rozmaitych problemów analitycznych. Na wstępie użytkownik wybiera rodzaj zagadnienia, jakie będzie analizowane. Można wybrać spośród następujących opcji:

- ◆ regresja (dla danych nietworzących szeregu czasowego),
- ◆ klasyfikacja (dla danych niesekwencyjnych),
- ◆ regresja szeregów czasowych,
- ◆ klasyfikacja szeregów czasowych,
- ◆ analiza skupień (z użyciem map Kohonena), czyli segmentacja.

Do najważniejszych funkcji programu *STATISTICA Automatyczne Sieci Neuronowe* należą m.in.: zintegrowane narzędzia do wstępnego przetwarzania danych i wyników, automatyczny projektant sieci, algorytmy uczące zbudowane na bazie najnowszych rozwiązań, w wysokim stopniu zoptymalizowane (w tym algorytm gradientów sprzężonych i BFGS), możliwość wyboru funkcji aktywacji, funkcji błędu i złożoności sieci, możliwość



konstruowania z sieci różnego typu zestawów sieci złożonych o praktycznie nieograniczonej wielkości, szeroki wybór narzędzi graficznych i statystycznych umożliwiających interakcyjną analizę eksploracyjną, pełna integracja z systemem STATISTICA.



Wszystkie wyniki, wykresy, raporty itp. mogą być przetwarzane za pomocą narzędzi graficznych i analitycznych STATISTICA (np. w celu analizy reszt z wyników prognoz, tworzenia podsumowujących raportów itd.). Program jest w pełni zintegrowany z narzędziami STATISTICA do automatyzacji analiz. Opcjonalnie dostępny jest generator kodu w języku C.

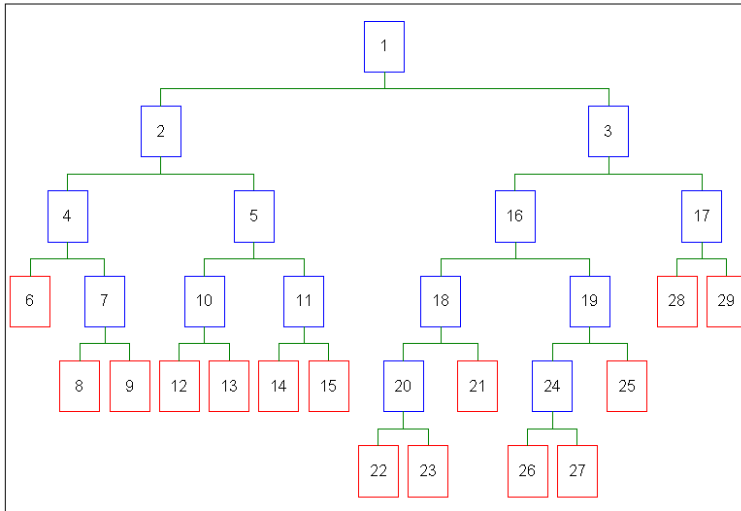


### STATISTICA Trees PL

Program STATISTICA Trees jest doskonały dla analityków chcących przeprowadzać analizy data mining przy wykorzystaniu modeli drzew decyzyjnych. W programie tym znajduje się część metod dostępnych także w STATISTICA Data Miner. Są to: *Ogólne modele drzew klasyfikacyjnych i regresyjnych (G-Trees)*, *Ogólne modele CHAID*, *drzewa interakcyjne i wzmacniane*. STATISTICA Trees umożliwia wykonywanie analiz z zastosowaniem algorytmu budowy drzew, które prowadzą do znalezienia zbioru logicznych warunków podziału typu *jeżeli, to*, pomagają jednoznacznie zaklasyfikować badane obiekty. Mechanizm eksploracji drzewa jest bardzo przydatny przy szczegółowej analizie wyników.



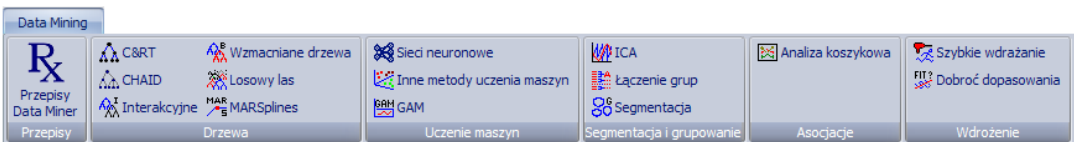
Układ drzewa 1 dla Zmn2  
Liczba węzłów dzielonych: 14, liczba węzłów końcowych: 15



### STATISTICA Data Miner PL

STATISTICA Data Miner to kompletny zestaw narzędzi data mining, zaprojektowany tak, aby umożliwić łatwe i szybkie wykonywanie analiz danych i stosowanie wyników we wspomaganiu podejmowania decyzji. STATISTICA Data Miner jest łatwym w użyciu, a jednocześnie bardzo wydajnym narzędziem do realizacji wszystkich etapów wydobywania z danych użytecznej wiedzy: począwszy od pobierania danych z baz i hurtowni danych, poprzez etap czyszczenia i transformacji danych, wykonywanie analiz danych, a na tworzeniu raportów skończywszy.

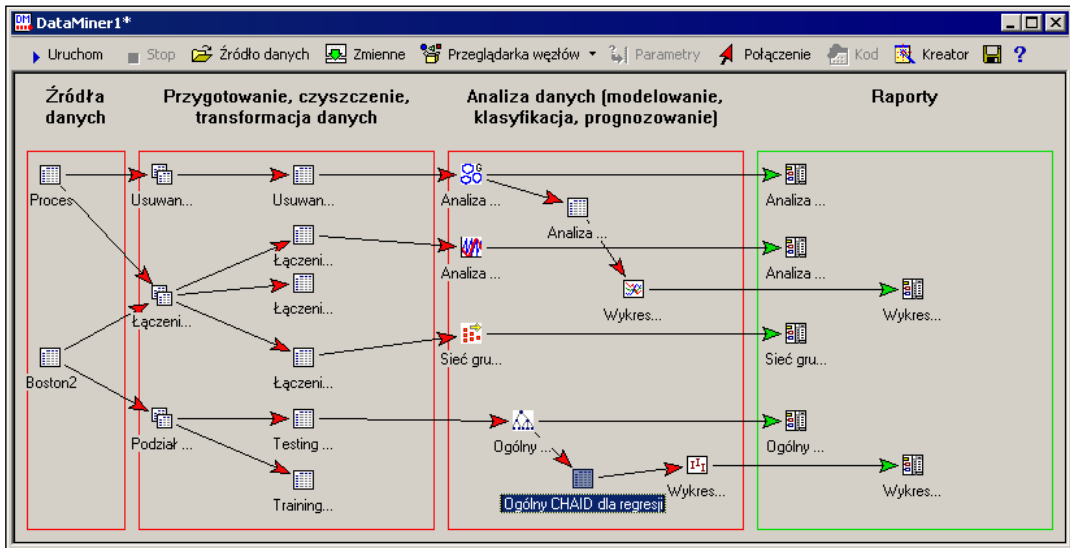
System STATISTICA Data Miner jest wyposażony w bardzo bogaty wybór metod analitycznych, zarówno wbudowanych w gotowe projekty data mining (szeroki wybór predefiniowanych, gotowych do użycia projektów dla różnego typu zagadnień), jak i dostępnych do budowy własnych projektów. STATISTICA Data Miner zawiera pełną funkcjonalność programów: STATISTICA Pakiet Podstawowy, Modele Zaawansowane, Analizy Wielowymiarowe, Automatyczne Sieci Neuronowe (z generatorem kodu C), Trees i Analiza Mocy Testów, składnik IDP, a przede wszystkim kilkanaście specjalnie zaprojektowanych modułów data mining.



Specjalne środowisko graficzne tworzenia projektów analitycznych (patrz poniżej) bazuje na konwencji *przeciagnij i upuść*, dzięki czemu tworzenie projektów data mining jest



intuicyjne, a praca z systemem nie wymaga długich przygotowań. Sprawia to, że z systemu *STATISTICA Data Miner* korzystać mogą nawet nowicjusze. Jest on jednocześnie wysoce elastyczny i dostosowywalny. Jak widać poniżej, obszar roboczy w *STATISTICA Data Miner* podzielony jest na cztery sekcje. Podział ten odpowiada czterem etapom projektów data mining.





















Dane przetwarzane w *STATISTICA* mogą pochodzić z wielu różnych źródeł. System *STATISTICA Data Miner* cechuje łatwość w korzystaniu z danych zewnętrznych. Mamy możliwość podłączenia różnych źródeł danych, tak że jakakolwiek zmiana danych na wejściu powoduje przeliczenie całego projektu, tak aby wyniki odpowiadały aktualnym danym. Wielkość danych przetwarzanych przez systemy *STATISTICA* jest praktycznie nieograniczona. Specjalne techniki data mining w *STATISTICA Data Miner* są zoptymalizowane pod kątem przetwarzania olbrzymich zbiorów danych.

Nowością w *STATISTICA Data Miner* są *Przepisy Data Miner* – kompletne rozwiązanie, dzięki któremu tworzenie modeli predykcyjnych staje się systematycznym procesem wykonywanym krok po kroku według ustalonej procedury. Celem *Przepisów Data Miner* jest umożliwienie łatwego i szybkiego wydobywania z danych użytecznych informacji. W *Przepisach* użytkownik tworzy model, wykonując krok po kroku pewien ogólny przepis obejmujący przygotowanie i oczyszczenie danych, a następnie użycie ich do stworzenia modeli predykcyjnych i zastosowanie modeli dla nowych obiektów. Wiele domyślnych ustawień zależy od danych, np. jeśli w pliku mamy wiele zmiennych, to musimy zastosować dobór zmiennych do modelowania, a jeśli zmiennych jest niewiele, to dobór zmiennych jest opcjonalny. Domyślne ustawienia systemu zazwyczaj dają zadowalające wyniki, ale zaawansowani użytkownicy mają do dyspozycji dodatkowe ustawienia i mogą zmieniać ustawienia domyślne.



*STATISTICA Data Miner* zawiera kilkanaście specjalnie zaprojektowanych modułów data mining:

-  Interakcyjne drażnienie danych
-  Ogólne modele drzew klasyfikacyjnych i regresyjnych (*G-Trees*)
-  Ogólne modele CHAID (*Chi-squared Automatic Interaction Detection*)
-  Drzewa interakcyjne (C&RT, CHAID)
-  Drzewa wzmacniane (*Boosted Trees*)
-  Automatyczne sieci neuronowe
-  Losowy las (*Random Forest*)
-  Analiza składowych niezależnych (ICA – *Independent Component Analysis*)
-  Analiza skupień uogólnioną metodą EM i k-średnich
-  Analiza koszykowa (*Association Rules, Basket Analysis*)
-  Uogólnione modele addytywne (*General Additive Models*)
-  MARSplines (*Multivariate Adaptive Regression Splines* – Wielozmienna regresja adaptacyjna z użyciem funkcji sklepanych)
-  Inne metody uczenia maszyn (*Machine Learning*): wektory nośne (*Support Vector Machines*), naiwny klasyfikator Bayesa, K-najbliższych sąsiadów (*K-Nearest Neighbors*).
-  Szybkie wdrażanie modeli predykcyjnych (PMML)
-  Dobroć dopasowania, klasyfikacji, predykcji (*Goodness of fit*)
-  Dobór i eliminacja zmiennych
-  Łączenie klas (grup) do predykcyjnego data mining
-  Przepisy Data Miner

Rozwiązania *STATISTICA* znajdują zastosowanie właściwie wszędzie tam, gdzie są gromadzone dane. Poniżej zamieszczono krótkie opisy pozostałych programów z rodziny *STATISTICA* wspomagających analizę danych w różnych dziedzinach zastosowań. Szczegółowe opisy wszystkich programów dostępne są na stronie internetowej [www.StatSoft.pl](http://www.StatSoft.pl).



***STATISTICA Enterprise*** jest zintegrowanym systemem analizy danych umożliwiającym współpracę analityków, współdzielenie danych, szablonów analiz i raportów.



***STATISTICA QC*** jest to zestaw statystycznych narzędzi zapewnienia i doskonalenia jakości, wykorzystywanych w przemyśle, badaniach naukowych, zapewnieniu jakości, rolnictwie, medycynie i wielu innych dziedzinach.



***STATISTICA MSPC*** (*Multivariate SPC* – *Wielowymiarowe SPC, procesy wsadowe*) to kompletne narzędzie do wielowymiarowego sterowania jakością procesów.



**STATISTICA VEPAC** (*Variance Estimation and Precision - szacowanie wariancji i precyzja*) to kompletny zestaw technik służący do analizy danych otrzymanych w wyniku doświadczeń, na które wpływ miały zarówno efekty stałe, jak i losowe.



**STATISTICA Process Optimization** (*QC Data Miner*) stanowi rozszerzenie systemu *STATISTICA Data Miner* o specjalistyczne procedury predykcyjnego data mining w zastosowaniach przemysłowych.



**STATISTICA MAS** (*Monitoring and Alerting Server*) umożliwia scentralizowane i zautomatyzowane monitorowanie wielu cech, zjawisk. Bieżące monitorowanie to automatyczna, efektywna metoda: a) równoczesnego monitorowania wielu krytycznych parametrów, b) dostarczania odpowiednich informacji o stanie badanych parametrów osobom o właściwych kompetencjach.



**STATISTICA Text Miner** jest rozszerzeniem systemu *STATISTICA Data Miner*, służącym do przekształcania pozbawionych struktury tekstów w użyteczną wiedzę.



**STATISTICA SAL** (*Sequence, Association and Link Analysis* – analiza sekwencji, asocjacji i połączeń) wykorzystuje szereg najnowszych technik analitycznych przeznaczonych do znajdowania w zbiorach danych reguł dotyczących wspólnego występowania zjawisk.



**STATISTICA Zestaw Farmaceutyczny** zawiera narzędzia wspomagające automatyzację analiz statystycznych w przemyśle farmaceutycznym. Składa się z programów: *Profile Uwalniania, Ocena Stabilności, Walidacja Procesu, Analiza Trendów, APR – Raporty Okresowe, Walidacja Metod Pomiarowych*.



**STATISTICA Zestaw Skoringowy** jest rozwiązaniem wspierającym proces budowy, oceny i utrzymania modeli zbudowanych za pomocą regresji logistycznej oraz innych metod predykcyjnych (np. SURVIVAL). Modele skoringowe zbudowane w *STATISTICA* wykorzystuje się np. przy wspomaganiu diagnostyki, określaniu prawdopodobieństwa występowania powikłań – wszędzie tam, gdzie wartości zmiennej dwustanowej (np. typu chory/zdrowy) chcemy objaśnić za pomocą innych zmiennych.



**STATISTICA Zestaw Laboratoryjny** zawiera narzędzia do automatycznej obróbki danych walidacyjnych oraz obróbki danych z badań i porównań międzylaboratoryjnych.



**STATISTICA w Six Sigma**. *STATISTICA* jest platformą analityczną w pełni wspierającą metodykę Six Sigma niezależnie od wielkości i zakresu jej implementacji.

**STATISTICA dla Analiz Marketingowych i Rynkowych** ułatwia i przyspiesza proces opracowania wyników badań, ze szczególnym uwzględnieniem badań ankietowych.

**Gromadzenie i analiza danych ankietowych**. Jest to rozwiązanie wspierające prowadzenie badań ankietowych, umożliwia m.in. wprowadzanie danych z uzupełnionych ankiet lub ich wypełnianie on-line, za pośrednictwem komputera z dostępem do Internetu i przeglądarkę.



## Usługi StatSoft Polska

W StatSoft myślimy o statystyce jako o nauce i sztuce zdobywania informacji z danych. Dlatego od wielu lat tworzone przez nas rozwiązania oraz oferowane usługi cieszą się uznaniem klientów. Zakres oferowanych przez nas usług jest bardzo szeroki począwszy od specjalistycznych kursów i szkoleń warsztatowych, poprzez kompleksowe usługi w zakresie projektowania i analizy wyników badań, aż po budowę systemów analitycznych dostosowanych do indywidualnych potrzeb użytkowników.

### *Szkolenia*

Istotnym elementem naszej oferty w zakresie kompleksowej analizy danych są specjalistyczne kursy i szkolenia, prowadzone przez znakomitych wykładowców, o bardzo bogatym doświadczeniu w zakresie zastosowań statystycznej analizy danych w medycynie. W naszej ofercie znajduje się bardzo obszerny zestaw szkoleń poświęconych różnym metodom analizy danych, w tym m.in. cały cykl szkoleń poświęconych wykorzystaniu metod statystycznych w medycynie i naukach biologicznych. Wszystkie nasze szkolenia mają charakter warsztatów, a omawiane metody ilustrowane są praktycznymi przykładami, ułatwiającymi zrozumienie idei poszczególnych metod i późniejsze ich zastosowanie, a szczególny nacisk położony jest na właściwą interpretację otrzymanych wyników.

### *Usługi konsultingowe*

Zakres świadczonych przez StatSoft Polska usług konsultingowych jest bardzo szeroki i obejmuje wszystkie obszary, w których analiza danych może przynieść wymierne korzyści w medycynie: od zaplanowania i przeprowadzenia badań, analizy ich wyników, aż po budowę zaawansowanych systemów automatycznie wykonujących analizy i raporty.

Nasi konsultanci mają szeroką wiedzę i duże doświadczenie w stosowaniu metod analizy danych w medycynie. Z powodzeniem realizowaliśmy wiele projektów dotyczących badania skuteczności metod leczenia, wspomagania diagnostyki medycznej, analizy przeżycia i poszukiwania czynników wpływających na przeżywalność, oceny jakości życia i wielu innych. Wynikiem takich projektów są liczne raporty oraz publikacje w renomowanych czasopiśmie. Do StatSoft Polska może się zatem zwrócić każdy, kto chciałby zasięgnąć porady w zakresie analizy danych lub chce skorzystać z pomocy naszych doświadczonych konsultantów.

## Licencje akademickie typu Site License

Mając na uwadze potrzeby instytucji naukowych i edukacyjnych, proponujemy takim jednostkom wygodny i powszechny dostęp do narzędzi analizy danych i data mining *STATISTICA*. Specjalne licencje akademickie typu Site License na programy i systemy z rodziny *STATISTICA* zapewniają wszystkim pracownikom i studentom danej jednostki



nieograniczony dostęp do najnowocześniejszych rozwiązań w zakresie statystycznej analizy danych, data mining oraz wizualizacji i graficznej analizy danych.

Najważniejsze zalety Licencji Akademickiej to m.in.:

- ◆ z programów i systemów objętych licencją może korzystać dowolna liczba użytkowników,
- ◆ w ramach licencji istnieje możliwość instalacji oprogramowania na komputerach domowych pracowników i studentów danej jednostki,
- ◆ w ramach jednej licencji programy mogą być instalowane zarówno w wersjach jedno-stanowiskowych, jak i sieciowych,
- ◆ licencja może obejmować zarówno polską, jak i angielską wersję językową,
- ◆ oprogramowanie może być wykorzystywane zarówno do celów dydaktycznych, jak i badań naukowych,
- ◆ w okresie obowiązywania licencji nowe, udoskonalone lub poprawione wersje oprogramowania są dostarczane automatycznie i bezpłatnie,
- ◆ użytkownikom licencji oferujemy specjalne warunki korzystania z usług szkoleniowych i konsultingowych oraz szeroki zakres pomocy technicznej,
- ◆ licencja może obejmować okresy jednoroczne lub dłuższe.

Oprócz Licencji Akademickiej dla wszystkich pracowników danej jednostki, dostępne są także licencje akademickie na określoną liczbę stanowisk komputerowych.

## StatSoft Polska

StatSoft Polska jest działającym od 1995 roku biurem amerykańskiej firmy StatSoft Inc., która ma już ponad dwudziestopięcioletnie doświadczenie w zakresie systemów analitycznych i związanych z nimi usług. Główną domeną działalności StatSoft Polska jest szeroko rozumiana analiza danych. Profesjonalne oprogramowanie analityczne *STATISTICA* oraz kompleksowe usługi doradcze i szkoleniowe realizowane przez doświadczonych specjalistów stanowią gwarancję zadowolenia klientów.

StatSoft Polska jest jednym z największych w Polsce dostawców programów do statystycznej analizy danych i data mining, a także największym w Polsce organizatorem specjalistycznych kursów i szkoleń dotyczących zastosowań analizy danych (statystycznej, data mining) w różnych dziedzinach. Członkowie zespołu StatSoft Polska brali udział w wielu projektach analitycznych i wdrożeniowych.

Rozwiązania StatSoft są powszechnie wykorzystywane ze względu na uniwersalność i nowoczesność zaimplementowanych metod, łatwość ich stosowania w specjalnie zaprojektowanym na potrzeby analizy danych interfejsie, bogatym możliwościom wizualizacji danych. W Polsce są to tysiące użytkowników w setkach firm, instytucji rządowych i samorządowych, uczelni, instytutów badawczych. Więcej informacji można znaleźć na stronie [www.StatSoft.pl](http://www.StatSoft.pl).