„Badanie współzależności atrybutów jakościowych w wielowymiarowych tabelach danych”

Przewodnik do ćwiczenia nr 1 dla studentów TPD w ramach przedmiotu „Zaawansowana eksploracja danych” (październik 2008)

Autorzy: *Jerzy Stefanowski*, *Robert Susmaga* Instytut Informatyki, Politechnika Poznańska.

Cel ćwiczenia: Celem jest wykorzystanie znanych miar oceny współzależności atrybutów do ustalania ważności poszczególnych atrybutów w wielowymiarowych tabelach danych. Pośrednio można ustalić nadmiarowe atrybutu w analizowanej tabeli i dokonać jej redukcji. Nabyta wiedza może być przydatna w dalszych zajęciach – Case 1 – do selekcji atrybutów w problemach klasyfikacyjnych.

Zakłada się, że atrybuty mogą być zdefiniowane na przynajmniej skali nominalnej. Do badania współzależności takich atrybutów można wykorzystać miary siły związku oparte na statystyce χ^2. Celem dydaktycznym jest wykorzystanie miary V- Cramera oraz poznanie oprogramowania Statsoft Statistica.

Wstępna wiedza: Studenci przed rozpoczęciem zajęć powinni przypomnieć wiedzę na temat wykonywania testu χ^2 (przedmiot Statystka i analiza danych na 3 roku studiów) oraz miar siły związku (poznać wzory na obliczanie miar Φ Yula i V – Cramera). Ponadto należy przypomnieć sobie zasady tworzenia tablic przestawnych w Excel oraz wykorzystywania funkcji statystycznych. W przypadku braku notatek z poprzedniego przedmiotu proszę skorzystać z książek nt. statystycznej analizy danych polecanych na wykładach.

Zadanie 1. Przypomnienie zasad badania zależności zmiennych nominalnych.
Wykonaj test niezależności dwóch zmiennych nominalnych i później oblicz współczynnik V-Cramera dla poniższej tabeli dwudzielczej (kontyngencji):

42	18
18	22

Uwaga: dla obliczeń testu przyjmij, że wartość krytyczna odpowiadająca założonemu poziomowi istotności wynosi 3.8 (Proszę zamieścić wszystkie wartości obliczeń częściowych. Odpowiedz na pytanie czy wartość współczynnika V-Cramera świadczy o dużej zależności pomiędzy badanymi zmiennymi?
Zadanie powyższe wykonaj na kartce papieru – wyniki zostaną przedyskutowane z prowadzącym.

Zadanie 2. Obliczenia z wykorzystaniem arkusza kalkulacyjnego Excel.
Powtórz wykonanie powyższego ćwiczenia w ramach funkcjonalności arkusza kalkulacyjnego Excel. Przypomnij sobie ograniczenia związane z użyciem funkcji TEST.CHI(). W jaki sposób dokonasz obliczenia współczynnika V-Cramera?

Zadanie 3. Analiza współzależności w wielowymiarowej tabeli danych oraz zapoznanie się z oprogramowanie Statistica.

Dostępne są pewne dane o klientach pozyskane w ramach badań marketingowych – patrz tabela zawarta w kolejnym pliku (marketingowe.xls) dostarczonym przez prowadzącego. Celem badania jest ustalenie zależności pomiędzy atrybutami charakteryzującymi klientów, w szczególności Płeć vs posiadanie samochodu, wielkość dochodu vs. posiadanie samochodu, wielkość dochodu vs. stan konta, itp.

Jeśli posługujesz się oprogramowaniem Statistica to wprowadź/zaimportuj omawiane dane (gdyby wystąpił błąd z importem xls utwórz nowe dane w Statistica o odpowiednich wymiarach i przekopiuj blok danych). Następnie stwórz odpowiednie tabele wielodzielcze (np. *Statsitica* - moduł „Podstawowe Statystyki”, opcja „tabele wielodzielcze”) i posługując się testem χ^2 (kolejne okno

wynik tabelaryzacji – pamiętaj o ustawieniu odpowiednich pól, tj. test Chi Pearsona i NW oraz wybierz odpowiedni poziom istotności α, zaznacz także miary Φ, V-Cramera) odpowiedz na pytanie:

Czy powyższe pary atrybutów/zmiennych są zależne?

Jeśli tak oceń siłę związku korzystając z dostępnych miar.

Zbuduj ranking najsilniejszych zależności między wartościami atrybutów, jakie odkryjesz w tej tabeli z danymi.

Zadanie 4. Analiza współzależności w wielowymiarowej tabeli danych (zoo data).

Dany jest zbiór danych zawierający 16 następujących charakterystyk zwierząt: *hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, tail, domesticated* and *catsize*. (wybierz właściwą zakładkę skoroszytu jedna to opis ang.. druga to same dane) Wszystkie atrybuty (z wyjątkiem atrybutu *legs*, którego wartością jest liczba nóg zwierzęcia) są nominalne. Interpretacja poszczególnych wartości jest następująca: wartości 1/0 atrybutu *hair* oznaczają: zwierzę jest/nie jest pokryte sierścią lub włosiem, wartości 1/0 atrybutu *feathers* oznaczają: zwierzę jest/nie jest pokryte piórami, itd.

W analogiczny sposób jak w poprzednim zadaniu zbadać statystyczną zależność pomiędzy wybranymi parami atrybutów i zinterpretować wyniki. Proszę odkryć najsilniejsze zależności dostępne w danych.

Zadanie 5. Analiza współzależności w wielowymiarowej tabeli danych (stress-level).

Ćwiczenie składa się z dwóch części:

W części pierwszej bada się zależności pomiędzy parami zmiennych, z których obie są dyskretne. Wykorzystywane pojęcia: tablice dwudzielcze (z j. ang. kontyngencji), dwuwymiarowy test χ^2, prawdopodobieństwo odrzucenia hipotezy zerowej, współczynnik V-Cramera.

W części drugiej bada się zależności pomiędzy parami zmiennych, z których jedna jest dyskretna, a druga ciągła (wymagana jest dodatkowa faza obliczeń). Wykorzystywane pojęcia: dyskretyzacja zmiennych ciągłych oraz pojęcia z części pierwszej.

Na dane składają się wyniki badań ankietowych, dostępne w pliku „stress-level;", które dotyczyły

- oddawania się wybranym rozrywkom (telewizja, internet) – dane dyskretne
- zażywania wybranych używek (kawa, papierosy, alkohol) – dane dyskretne
- poziomu odczuwanego stresu – dana zdefiniowana na skali liczbowej.

Dostępny zbiór danych przedstawia zebrane odpowiedzi 300 ankietowanych (wiersze) opisane w kategoriach 6 zmiennych (kolumny) – czyli tworzy tablicę wielokolumnową. Wszystkie zmienne oprócz zmiennej zawierającej informacje o poziomie odczuwanego stresu są zmiennymi dyskretnymi.

Interesującymi badawczo pytania są następujące:

Czy można stwierdzić istnienie statystycznie istotnych zależności pomiędzy intensywnością zażywanych rozrywek?

Czy można stwierdzić istnienie statystycznie istotnych zależności pomiędzy intensywnością zażywanych używek?

Czy można stwierdzić istnienie statystycznie istotnych zależności pomiędzy intensywnością zażywanych rozrywek i używek? Które z tych zależności są najsilniejsze?

Czy można stwierdzić istnienie statystycznie istotnych zależności pomiędzy intensywnością zażywanych rozrywek/używek a poziomem odczuwanego stresu, a jeżeli tak, to z jakimi rozrywkami/używkami poziom ten jest najbardziej związany?

Proponowany przebieg ćwiczenia

Zapoznanie się z zawartością pliku z wielowymiarowymi danymi.

Część pierwsza

1) Tworzenie tablic dwudzielczej (kontyngencji) dla zmiennych dyskretnych.

2) Wprowadzenie wartości oczekiwanych i testu zależności dla zmiennych nominalnych opartego na statystyce χ^2 (tzw. dwuwymiarowy test χ^2).

3) Testowanie zależności pomiędzy wybranymi parami zmiennych dyskretnych w oparciu o test χ^2.

4) Ocena siły związku między wybranymi parami zmiennych dyskretnych w oparciu o wartość współczynnika V-Cramera.

Podsumowaniem powinno być stworzenie rankingu par zmiennych pod względem siły związku, ustalenie czy któreś z par zmiennych nie są w relacji współzmienności oraz interpretacja otrzymanych wyników.

Część druga

1) Analiza zmiennej liczbowej.

2) Dyskretyzowanie tej zmiennej (przeanalizuj wartości tej zmiennej i zaproponuj najdogodniejszą metodę dyskretyzacji dziedziny tej zmiennej).

3) Testowanie zależności pomiędzy zdyskretyzowaną zmienną a wybranymi zmiennymi dyskretnymi (jak w części pierwszej) oraz ustalenie, które ze zmiennych (rozrywki, używki) mają najsilniejszy wpływ na wartości zmiennej „poziom stresu”.