Středoškolská odborná činnost

Obor 01 - Matematika a statistika

Aplikace diskriminační analýzy

Autoři:	Ondřej Ficker
	Petr Langer
	Robert Stárek
Škola:	Slovanské gymnázium Olomouc,
	tř. Jiřího z Poděbrad 13
	Olomouc, 771 10
Studíjní obor:	všeobecný se zaměřením na přírodní vědy
	3. ročník
Konzultant:	RNDr. Karel Hron, Ph.D.
	Přírodovědecká fakulta UP Olomouc,
	Katedra matematické analýzy a aplikací matematiky

Olomouc, 2009

Prohlášení:
Prohlašujeme, že jsme naši práci vypracovali samostatně a použili jsme pouze podklady (literaturu, SW atd.) uvedené v přiloženém seznamu. Nemáme závažný důvod proti zpřístupňování této práce v souladu se zákonem č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) v platném znění.

V Olomouci dne 25.3.2009 Podpisy:

Poděkování:
Chtěli bychom poděkovat prof. Hronovi za trpělivé vysvětlování tajemství královny věd nám, obyčejným smrtelníkům. Za jeho pomoc při hledání vhodné aplikace a především za jeho trpělivost při neustálých drobných formálních opravách naší práce. Dále děkujeme všem lidem, kteří byli ochotni poskytnout nám údaje, na kterých byl náš výzkum postaven, bez takto ochotných by neměl podobný výzkum smysl. Vážíme si také ochoty našich rodičů a sourozenců pomoci s hledáním aplikace a zhodnotit naši práci z jejich pohledu. Děkujeme také našim profesorům za uvolnění z jejich hodin z důvodu práce na této soč.

Obsah

Úvod

Statistika je výborným nástrojem všech věd, ať už přírodních, nebo humanitních. Ať už se rozhodneme pro studium jakékoli vysoké školy, je jisté, že se s touto disciplínou ve větší či menší míře setkáme. Vždy je potřeba získat určitou představu o vlastnostech množiny daných objektů. Může se přitom jednat o elementární částice, jedince určitého živočišného druhu nebo třeba anglická slova. Zkrátka všude, kde je nutné pracovat s velkými objemy dat, je potřeba dívat se na soubor objektů jako na celek a zkoumat charakteristické znaky, jejich vzájemné vztahy, popřípadě objekty podle těchto znaků charakterizovat. A v poslední době těchto možností uplatnění statistiky rapidně přibývá. Všechny tyto argumenty nás přesvědčily, abychom se použití statistiky věnovali i v naší práci.

Jednou z velmi zajímavých metod, které slouží právě ke kategorizaci statistických objektů, je diskriminační analýza. Má sice už teď poměrne široké spektrum využití, ale stále v některých odvětvích nenašla uplatnění, a tak nás zajímalo, kde všude by se tato metoda ještě dala použít. Velmi slibná je možnost uplatnění diskriminační analýzy v medicíně, ovšem vzhledem k ohromné složitosti a variabilitě lidského organismu je nutné diskriminační analýzu správně interpretovat, aby byla v lékařské praxi využitelná. Jednoduchý příklad aplikace diskriminační analýzy v medicíně uvádíme v naší práci. Dalším možným uplatněním, které nás v souvislosti s diskriminační analýzou napadlo, je uplatnění v sociologii a marketingu. Tomuto perspektivnímu uplatnění jsme se věnujeme především. S pomocí lidí v našem okolí se nám podařilo sestavit charakteristiku typického člověka, který (ne)používá internetové služby typu ICQ a nalézt rozhodovací kritérium pro zařazení nového objektu.

Samozřejmě, že bylo zapotřebí k naší práci nastudovat a pochopit řadu matematických pojmů a metod. I to bylo naší motivací - vždy je dobré naučit se něco nového, zvlášť pokud se jedná o široce uplatnitelná fakta a metody. Je obvyklé, že lidská zvědavost dokáže přinutit i lenochy, aby se začali prokousávat mnohdy složitou literaturou, nejinak tomu bylo u nás. Doufáme, že

nám naše bádání nejen rozšíří znalosti i dovednosti v oboru matematiky, ale že nám přinese i zajímavé výsledky našich měření.

1 Základy maticové algebry

Abychom mohli efektivně pracovat s pojmy, zavedenými později v diskriminační analýze, zmíníme se na úvod o základních poznatcích z maticové algebry. Matic se využívá v matematice, fyzice a především ve statistice pro práci s velkými objemy dat,neboli rozsáhlými statistickými soubory. Matice, které jsou v jistém smyslu zobecněním reálného čísla, můžeme stejně jako běžná čísla sčítat či násobit jiným číslem, násobení dvou matic se již řídí zvláštními pravidly, [5],[3].

1.1 Základní druhy matic

Definice 1 Nechť jsou dána přirozená čísla $m, n \in N$ a reálná čísla $a_{ij} \in R$, $i = 1, \ldots, n, j = 1, \ldots, m$. Blokové schéma o m řádcích a n sloupcích nazveme maticí o rozměrech $m \times n$,

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{n2} & \ldots & a_{mn} \end{pmatrix} = (a_{ij})_{i,j=1}^{m,n}.$$

Matice běžně značíme velkými tučnými písmeny. Reálné číslo je z tohoto pohledu matice o rozměrech 1×1. Vektor je matice o rozměrech $n \times 1$, budeme tedy uvažovat vektory sloupcové. Pro naše účely je vhodné definovat nejprve některé speciální typy matic.

Definice 2 Matici nazýváme *čtvercovou maticí řádu n*, je-li počet jejích řádků m rovný počtu jejich sloupců n. Matice $\mathbf{A}$ pak bude ve tvaru

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}.$$

Příklad 1 Čtvercová matice řádu 3 může vypadat třeba takto:

$$\mathbf{I}_3 = \begin{pmatrix} \pi & -4 & 3 \\ 0 & 8 & -6 \\ -1 & \sqrt{3} & 1 \end{pmatrix}.$$

Diagonálou čtvercové matice $\mathbf{A}$ pak nazveme n-tici reálných čísel $a_{11}, \dots, a_{nn}$.

Definice 3 Matici nazýváme *jednotkovou maticí řádu n*, je-li čtvercová řádu n a má-li na hlavní diagonále jedničky a na ostatních pozicích nuly. Jednotkovou matici značíme $\mathbf{I}_n$.

Příklad 2 Čtvercová matice řádu 3 bude ve tvaru

$$\mathbf{I}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Definice 4 Maticí *transponovanou* k matici $\mathbf{A} = (a_{ij})_{i,j=1}^{m,n}$ nazveme matici

$$\mathbf{A}' = (a_{ji})_{j,i=1}^{n,m}.$$

Transpozici matice provedeme záměnou dolních indexů prvků matice.
Prvek z i-tého řádku a j-tého sloupce tedy přesuneme do j-tého řádku a i-tého sloupce, vlastně tak zaměníme řádky a sloupce matice $\mathbf{A}$.

Příklad 3 Pro matici $\mathbf{A}$ o rozměrech 2×3, která je zadána:

$$\mathbf{A} = \begin{pmatrix} 1 & \sqrt{2} \\ 3 & -2 \\ 0 & 6 \end{pmatrix}$$

po transpozici obdržíme

$$\mathbf{A}' = \begin{pmatrix} 1 & 3 & 0 \\ \sqrt{2} & -2 & 6 \end{pmatrix}.$$

Definice 5 *Permutací* číselné množiny $A = \{1, 2, ..., n\}$ rozumíme zobrazení prvků této množiny na sebe, tedy když každému jejímu prvku přiřadíme opět prvek množiny A tak, že každý z nich bude přiřazen právě jednou. Permutaci zapisujeme jako dvouřádkové schéma,

$$\begin{pmatrix} 1 & 2 & 3 & ... & n \\ i_1 & i_2 & i_3 & ... & i_n \end{pmatrix},$$

které značí přiřazení $1 \to i_1, 2 \to i_2, \ldots, n \to i_n$, kde $1 \leq i_j \leq n$,
$j = 1, ..., n$ [5].

Příklad 4 Permutace P množiny $A = \{1, 2, 3, 4\}$ je

$$P = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \end{pmatrix}.$$

Definice 6 Říkáme, že i_j, i_k, kde $1 \leq j \leq n$, $1 \leq k \leq n$ a $j \neq k$, stojí v inverzi, jestliže platí $i_j \leq i_k$ a přitom $j > k$. Číslo $sgn(P) = (-1)^p$, kde p je počet inverzí v permutaci P, pak nazýváme znaménko permutace.

Příklad 5 Známenko permutace P uvedené v příkladu 4 je

$$sgn(P) = (-1)^3 = -1,$$

protože v této permutaci tvoří inverze právě dvojice prvků $3,1; 3,2; 4,2$.

Definice 7 Nechť $\mathbf{A} = (a_{i,j})_{ij=1}^{n,n}$ je čtvercová matice řádu n. Potom číslo $det(\mathbf{A}) = \sum_p sgn(P) a_{1p(1)} ... a_{np(n)}$, kde součet probíhá přes všechny permutace množiny $\{1, 2, ..., n\}$, nazýváme *determinantem* matice $\mathbf{A}$. Determinant matice $\mathbf{A}$ značíme též $|\mathbf{A}|$ [5].

Determinant čtvercové matice $\mathbf{A}$ vypočteme ve speciálních případech $n = 2$ a $n = 3$ snadno, pomocí tzv. Cramerova pravidla. K výpočtu determinantů matic vyšších řádů je užívána výpočetní technika.
Přitom platí:
pro $n = 2$

$$\begin{vmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21},$$

pro $n = 3$

$$\begin{vmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \end{vmatrix} = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \quad (1)$$

$$-(a_{13}a_{22}a_{31} + a_{12}a_{21}a_{33} + a_{11}a_{23}a_{32}). \quad (2)$$

Definice 8 Čtvercovou matici nazýváme *regulární*, jestliže její determinant je různý od nuly, jinak ji nazýváme *singulární*. Pokud je matice regulární, je možné spočítat k ní inverzní matici.

1.2 Početní operace s maticemi

Sčítání dvou matic

Definice 9 Mějme matice $\mathbf{A}$, $\mathbf{B}$ o stejných rozměrech, potom jejich součtem nazveme matici $\mathbf{A}$+$\mathbf{B}$, jejíž prvky jsou rovny součtu prvků matic $\mathbf{A}$,$\mathbf{B}$ na odpovídajících si pozicích, tedy pro matice

$$\mathbf{A} = (a_{ij})_{i,j=1}^{m,n} \quad \mathbf{B} = (b_{ij})_{i,j=1}^{m,n}$$

platí

$$\mathbf{A}_{ij} + \mathbf{B}_{ij} = (a_{ij} + b_{ij})_{i,j}^{/} m, n.$$

Pro sčítání matic platí zákon komutativní a zákon asociativní, tj. pro vhodně zvolené matice $\mathbf{A}, \mathbf{B}, \mathbf{C}$ máme

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}.$$

$$(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$$

Příklad 6 Situaci si objasníme pro matice $\mathbf{A}$ a $\mathbf{B}$ o rozměrech 3×2,

$$\mathbf{A} = \begin{pmatrix} -3 & 4 \\ 6 & 0 \\ \sqrt{5} & -2 \end{pmatrix},$$

$$\mathbf{B} = \begin{pmatrix} \pi & -2 \\ 3 & -1 \\ 1 & 6 \end{pmatrix};$$

pak bude součet těchto matic vypadat takto:

$$\mathbf{A} + \mathbf{B} = \begin{pmatrix} -3 & 4 \\ 6 & 0 \\ \sqrt{5} & -2 \end{pmatrix} + \begin{pmatrix} \pi & -2 \\ 3 & -1 \\ 1 & 6 \end{pmatrix} = \begin{pmatrix} -3+\pi & 2 \\ 9 & -1 \\ 1+\sqrt{5} & 4 \end{pmatrix}.$$

Násobení matice skalárem

Definice 10 Nechť je dána matice $\mathbf{A} = (a_{ij})_{i,j=1}^{m,n}$a reálné číslo c, potom c-*násobkem matice*$\mathbf{A}$ nazveme matici $c\mathbf{A}$ pro jejíž prvky platí $c\mathbf{A} = (c \cdot a_{ij})_{i,j=1}^{m,n}$ Násobení matice číslem (skalárem),tedy provádíme vynásobením všech prvků matice tímto číslem. Pro násobení matice skalárem (nebo více skaláry) platí zákon komutativní i asociativní, tedy

$$c\left(\mathbf{A}+\mathbf{B}\right) = c\mathbf{A} + c\mathbf{B}$$
$$\left(c+d\right)\mathbf{A} = c\mathbf{A} + d\mathbf{A}$$
$$\left(cd\right)\mathbf{A} = c\left(d\mathbf{A}\right)$$
$$1\mathbf{A} = \mathbf{A}.$$

Příklad 7 Násobením matice

$$\mathbf{A} = \begin{pmatrix} 2 & -4 \\ -6 & 8 \\ -10 & 12 \end{pmatrix}$$

číslem $c = 2$ dostaneme:

$$c\mathbf{A} = 2 \cdot \begin{pmatrix} 2 & -4 \\ -6 & 8 \\ -10 & 12 \end{pmatrix} = \begin{pmatrix} 4 & -8 \\ -12 & 16 \\ -20 & 24 \end{pmatrix}$$

Násobení dvou matic

Definice 11 Nechť je dána matice $\mathbf{A} = (a_{ij})_{i,j=1}^{m,n}$ o rozměrech $m \times n$ a matice $\mathbf{B} = (b_{jk})_{j,k=1}^{n,p}$ o rozměrech $n \times p$. Potom součinem matic $\mathbf{A}, \mathbf{B}$ (v tomto pořadí) nazveme $\mathbf{C} = \mathbf{AB} = (c_{ik})_{i,k=1}^{m,p}$, kde

$$c_{ik} = \sum_{j=1}^{n} a_{ij} b_{jk}.$$

Z definice vyplývá, že abychom mohli násobit dvě matice mezi sebou, musí mít první z nich stejný počet sloupců, jako má druhá matice řádků. Násobení matic je asociativní, ale není komutativní, tj. obecně neplatí rovnost $\mathbf{AB} = \mathbf{BA}$. Vlastně se jedná o skalární součin vektoru příslušného řádku první matice s vektorem sloupce druhé matice. Tento výsledek se pak zapíše na pozici ve výsledné matici $\mathbf{C}$, jejíž index odpovídá číslu řádku první matice a číslu sloupce druhé matice.

Příklad 8 Násobením matice $\mathbf{A}$ o rozměrech 2×3 maticí $\mathbf{B}$ o rozměrech 3×4

$$\mathbf{A} = \begin{pmatrix} 2 & -1 & 0 \\ 3 & 1 & 4 \end{pmatrix},$$

$$\mathbf{B} = \begin{pmatrix} 2 & 4 & -1 & -7 \\ 6 & 8 & 3 & -3 \\ 0 & 2 & 5 & 0 \end{pmatrix}$$

dostaneme matici

$$\mathbf{C} = \begin{pmatrix} 2 & -1 & 0 \\ 3 & 1 & 4 \end{pmatrix} \cdot \begin{pmatrix} 2 & 4 & -1 & -7 \\ 6 & 8 & 3 & -3 \\ 0 & 2 & 5 & 0 \end{pmatrix} = \begin{pmatrix} -2 & 0 & -5 & 11 \\ 12 & 28 & 20 & -24 \end{pmatrix}.$$

Přitom v tomto případě součin $\mathbf{BA}$ neexistuje, protože počet sloupců matice $\mathbf{B}$ je roven čtyřem a matice $\mathbf{A}$ má pouze dva řádky.

Definice 12 Pro *inverzní* matici k regulární čtvercové matici $\mathbf{A}$ řádu n, kterou značíme $\mathbf{A}^{-1}$, platí, že když ji vynásobíme původní regulární maticí, dostaneme jednotkovou matici daného řádu, tedy

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{I}_n.$$

Výpočet inverzní matice provádíme pomocí Gaussovy eliminační metody.
Na levou stranu soustavy umístíme původní matici, na pravou stranu příslušnou jednotkovou matici [5]. Upravujeme (pomocí povolených úprav - tzn. sčítání řádků, násobení řádků číslem a záměny pozic řádků) dokud nedostaneme nalevo jednotkovou matici a napravo inverzní matici, symbolicky

$$(\mathbf{A}|\mathbf{I}_n) \sim (\mathbf{I}_n|\mathbf{A}^{-1}).$$

2 Úvod do popisné statistiky

Popisná statistika poskytuje číselné i jiné údaje (např. grafické) o hromadných jevech ve všech oblastech společnosti. My se budeme soustředit zejména na pojmy, které budou důležité pro samotnou diskriminační analýzu. Výchozím pojmem popisné statistiky je *statistický soubor*, tedy definovaná množina *statistických jednotek*. Statistický soubor může být určen seznamem svých prvků (jednotek, např. osoby) nebo může být definován nějakým pravidlem, které slouží k selekci těchto prvků. U každé statistické jednotky měříme hodnotu nějakého *statistického znaku* X, nebo více takových statistických znaků, příkladem může být výška osob v cm a jejich váha v kg. Číselné chrakteristiky statistického znaku X budeme rozdělovat na dva typy - *charakteristiky polohy* a *charakteristiky proměnlivosti* (variability),[2].

Definice 13 Základní charakteristikou polohy je mimo jiné *aritmetický průměr* souboru hodnot $x_1, ..., x_n$ statistického znaku X, který definujeme jako

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Poznamenejme ovšem, že aritmetický průměr lze charakterizovat jako číslo, kolem kterého se hodnoty statistického znaku nejvíce pohybují. Aritmetický průměr je velmi citlivý na přítomnost odlehlých hodnot v souboru. Příkladem je situace, kdy nám jedna pětka zkazí jinak pěkný průměr známek.

Definice 14 Proměnlivost charakterizuje *rozptyl statistického znaku X*, který vyjadřuje variabilitu hodnot statistického znaku kolem aritmetického průměru. Nazveme jím číslo

$$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_1 - \overline{x})^2.$$

Řekněme si něco o geometrické interpretaci rozptylu. Z tohoto pohledu je rozptyl roven aritmetickému průměru obsahů čtverců o stranách $x_i - \bar{x}$, kde $i = 1, ..., n$. Musíme být obezřetní při počítání rozptylu z dat, které obsahují i odlehlé hodnoty (značně vzdálené od aritmetického průměru), protože nám tyto mohou rozptyl neúměrně zvýšit.

Definice 15 Z hlediska interpretace je někdy místo rozptylu výhodné používat jeho druhou odmocninu, která se nazývá *směrodatná odchylka*, tedy

$$s_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Pro dvojici statistických znaků X a Y zavádíme navíc další charakteristiku - tzv. *kovarianci*, která vyjadřuje lineární vztah mezi znaky X a Y.

Definice 16 Kovariancí dvou statistických znaků X a Y nazýváme číslo s_{XY},

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Poznamenejme, že pro vyjádření vztahů mezi statistickými znaky se často používá také bezrozměrná obdoba kovariance, tzv. *korelační koeficient* znaků X a Y, který je definován jako

$$r_{XY} = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}} = \frac{s_{XY}}{s_X s_Y}.$$

Korelačním koeficientem se však nebudeme blíže zabývat.

Jak kovariance, tak korelační koeficient jsou ze své definice symetrické, tzn. nezáleží na tom, zda počítáme kovarianci mezi X a Y, nebo Y a X, tedy $s_{XY} = s_{YX}$ Pro statistické znaky X a Y se čtveřice $s_X^2, s_Y^2, s_{XY}, s_{YX}$ často souhrnně nazývá *kovarianční struktura znaků X a Y.*

Varianční matice

Pro p-tici statistických znaků $X_1, \cdots, X_p$, kde $p \geq 2$, zapisujeme jejich kovarianční strukturu souhrnně pomocí tzv. *varianční matice.*

Definice 17 *Varianční maticí* statistických znaků $X_1, \cdots, X_p$, uspořádaných do vektoru $\mathbf{X} = (X_1, ...X_p)'$, nazýváme čtvercovou matici jejich rozptylů a kovariancí, tj. matici ve tvaru

$$\mathbf{S_X} = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 \end{pmatrix}.$$

Pokud budeme v dalším textu uvažovat více než dva statistické znaky, místo X a Y je budeme vždy označovat $X_1, X_2, \cdots, X_n$ a u jejich číselných charakteristik budeme uvádět pouze příslušný dolní index.

Na jednoduché sitauci si ukážeme výpočet jednotlivých číselných charakteristik pro dva statistické znaky X a Y a sestavíme varianční matici.

Příklad 9 Nechť X je výška studenta v centimetrech a Y vyjadřuje jeho váhu v kilogramech. Měřením byly obdrženy tyto hodnoty:

znak X	165	173	190	181
znak Y	55	78	87	81

Nejprve vypočítáme aritmetický průměr pro znak X,

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\overline{x} = \frac{1}{4}(165 + 173 + 190 + 181) = 177{,}25;$$

poté zcela analogicky pro znak Y,

$$\overline{y} = \frac{1}{4}(55 + 78 + 87 + 81) = 75{,}25.$$

Obdržíme vektor průměrů,

$$\begin{pmatrix} 177{,}25 \\ 75{,}25 \end{pmatrix}.$$

Dále spočítáme rozptyly

$$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2, s_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_1 - \overline{y})^2,$$

$$s_X^2 = 114{,}92, s_Y^2 = 196{,}25,$$

a nakonec kovariance

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = 135{,}92.$$

Výsledná varianční matice $\mathbf{S}_{(X,Y)'}$ pak bude ve tvaru

$$\mathbf{S}_{(X,Y)'} = \begin{pmatrix} 114,62 & 135,92 \\ 135,92 & 196,25 \end{pmatrix}.$$

3 Diskriminační analýza

Uvažujme nyní situaci, kdy máme dáno k různých skupin nějakých objektů, kde $k \geq 2$. Pod těmito skupinami a objekty rozumíme například pacienty, roztříděné dle závažnosti jejich onemocnění. Každý objekt je charakterizován p-ticí hodnot $\mathbf{x} = (x_1, ..., x_p)'$, tedy realizací statistických znaků $X_1, ..., X_p$. Předpokládáme, že na základě nějaké dřívější zkušenosti o rozdělení objektů do tříd jsme pro jednotlivé třídy určili vektory průměrů $\bar{\mathbf{x}}_i$ a varianční matice $\mathbf{S}_i$, $i = 1, \ldots, k$. Potom je zkoumán objekt, o němž není známo, do které

skupiny patří, ale přitom víme, že do jedné z nich patřit musí. Lze pouze zjistit hodnoty statistických znaků $X_1, ..., X_p$, které mu odpovídají. Na základě těchto zjištěných hodnot se má rozhodnout, do které z k skupin zadaný objekt patří. V naší situaci tedy půjde o to odhadnout, jaký průběh bude mít onemocnění nově příchozího pacienta.

Původní motivací pro vznik diskriminační analýzy byly problémy z oblasti antropologie [1]. Při vykopávkách mohli antropologové podle nalezených předmětů určit, ke kterému stadiu vývoje pravěkého člověka patří nalezené kostry. Tyto kostry mohli také podrobně proměřit a zaznamenat délky různých kostí či úhly, které kosti svíraly. Jestliže na jiném místě byla nalezena další kostra bez identifikačních předmětů, vznikl problém, ke kterému vývojovému stadiu ji zařadit jen na základě jejích charakteristik. V současné době se však s aplikacemi diskriminační analýzy setkáváme v nejrůznějších vědních oborech.

3.1 Formulace problému

Vraťme se ovšem nyní k původní formulaci problému. Rozhodování musíme založit na realizaci vektoru statistických znaků $\mathbf{X} = (X_1, \ldots, X_p)$, která nabývá hodnot z množiny všech p-rozměrných vektorů, kterou označíme R^p. Tedy v R^p zvolíme k podmnožin $A_1, \ldots, A_k$, které se nepřekrývají a jejichž sjednocení dává celý prostor R^p. Stručně říkáme, že tyto množiny tvoří rozklad R_p a z našeho pohledu vlastně tvoří množiny hodnot pro jednotlivé skupiny.

Je třeba navrhnout takový rozklad, aby rozhodnutí bylo optimální, tzn. aby pravděpodobnost zařazení do chybné skupiny byla co nejmenší. Úvahy o tomto optimálním rozkladu ovšem hluboko překračují naše matematické znalosti, přičemž se opírají o pojmy z teorie pravděpodobnosti, mimo jiné o tzv. mnohorozměrné normální rozdělení pravděpodobnosti. Proto se nyní zaměříme až na konkrétní rozhodovací pravidlo, [1].

Princip je následující. Označme si D_j tzv. *diskriminační* skóre pro j-tou skupinu, $j = 1, \ldots, k$. Potom objekt charakterizovaný vektorem $\mathbf{x}$ zařadíme do j-té skupiny, jestliže číslo $D_j > D_i$, kde $j \neq i$. V praxi tedy postupujeme tak, že pro vektor x vypočteme hodnoty $D_1, \ldots, D_k$ a poté objekt zařadíme do té

skupiny, která odpovídá největší z těchto hodnot.

Problémem je určit, s jakou pravděpodobností π_j patří objekt do j-té skupiny. Většinou se tato pravděpodobnost volí jako podíl objektů, které byly do této skupiny zařazeny na základě dřívější zkušenosti. Dále, jestliže je o některých objektech předem známo, do kterých skupin patří, provede se pro ně výpočet hodnot D_j a zjistí se, kam by který z nich byl zařazen. Získá se tak alespoň hrubá představa o účinnosti diskriminační metody, týkající se správného zařazení objektů do skupin, tedy o podílu správně zařazených objektů a celkového počtu zkoumaných objektů. Tuto charakteristiku pak vezmeme v úvahu při zařazování objektů, u kterých už není předem známo, do jaké skupiny patří.

3.2 Výpočet diskriminačního skóre

Nyní si uvedeme kompletní vzorec pro výpočet diskriminačního skóre. V našem případě se jedná o tzv. kvadratickou diskriminaci, protože každá ze skupin je charakterizována příslušnou varianční maticí S_j. U tzv. lineární diskriminace, která se též někdy užívá, jsou všechny skupiny charakterizovány stejnou varianční maticí a liší se tedy pouze vektory průměrů statistických znaků $\mathbf{x}_j, j = 1, \ldots, k$. Kvadratická diskriminace je proto výrazně přesnější metodou a více odráží reálnou situaci. Pro diskriminační skóre D_j j-té skupiny tedy platí

$$D_j = -\frac{1}{2} \ln |\mathbf{S}_j| - \frac{1}{2} (\mathbf{x} - \overline{\mathbf{x}}_j)' \mathbf{S}_j^{-1} (\mathbf{x} - \overline{\mathbf{x}}_j) + \ln \pi_j,$$

kde $\mathbf{S}_j$ je příslušná varianční matice daných statistických znaků a π_j je pravděpodobnost, že objekt bude zařazen do dané skupiny. Objekt $\mathbf{x}$ pak zařadíme do skupiny, pro kterou je D_j největší.

Příklad 10 [1] Uveďme situaci, kdy do nemocnice přicházejí pacienti s vážnou chorobou ihned po jejím vypuknutí. Podle dalšího průběhu choroby lze pacienty rozdělit do tří skupin:

1. Pacienti, kteří na tuto chorobu zemřeli do deseti dnů.

2. Pacienti, kteří na tuto chorobu umřeli po deseti dnech, nebo zůstali s trvalými následky.

3. Pacienti, kteří se zcela vyléčili.

Po přijetí pacienta lze jeho stav charakterizovat na základě anamnézy statistickým znakem X_1 a na základě předběžného lékařského vyšetření veličinou X_2. Až dosud bylo přijato 400 pacientů.

j-tá skupina	1	2	3
Počet pacientů	50	50	300
Vektor průměrů $\mathbf{x}_j$	$(5;5)'$	$(4;8)'$	$(6;9)'$
Varianční matice $\mathbf{S}_j$	$\begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$	$\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$

Nově přijatý pacient je charakterizován hodnotami $(5;7)'$. Je třeba již na základě těchto údajů rozhodnout do které ze tří skupin nejspíše patří. Máme

$$\pi_1 = \frac{50}{400} = \frac{5}{8}, \quad \pi_2 = \frac{50}{400} = \frac{1}{8}, \quad \pi_3 - \frac{300}{400} = \frac{3}{4}.$$

Spočítáme determinant varianční matice $\mathbf{S}_j$ pro první skupinu,

$$|\mathbf{S}_1| = 3 \cdot 2 - 1 \cdot 1 = 6 - 1 = 5.$$

Dále spočítáme inverzní matici k v $\mathbf{S}_1$. U čtvercové matice řádu 2 stačí zaměnit prvky na hlavní diagonále, zaměnit znaménka nediagonálních prvků a výslednou matici pak vydělit determinntem původní matice. Takto obdržíme

$$\begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}^{-1} = \frac{1}{5}\begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix} = \begin{pmatrix} \frac{2}{5} & -\frac{1}{5} \\ -\frac{1}{5} & \frac{3}{5} \end{pmatrix}.$$

Nyní stačí jen dosadit do příslušného vzorce,

$$D_1 = -\frac{1}{2}\ln 5 - \frac{1}{2}\cdot(2\ \ 0)\cdot\begin{pmatrix} \frac{2}{5} & -\frac{1}{5} \\ -\frac{1}{5} & \frac{3}{5} \end{pmatrix}\cdot\begin{pmatrix} 2 \\ 0 \end{pmatrix} + \ln\frac{5}{8} \quad (3)$$

$$= \ln 5 - \frac{1}{2}\cdot\frac{8}{5} + \ln\frac{5}{8} = -4{,}08 \quad (4)$$

Dále analogicky též pro ostatní skupiny, dohromady tedy

$$D_1 = -4{,}08 \ \ D_2 = -4{,}58 \ \ D_3 = -3{,}17.$$

Protože D_3 je maximální, budeme počítat s tím, že nový pacient nejspíše patří do třetí skupiny.

4 Statistický software R

Pro výpočty, nutné pro diskriminační analýzu, jsme využili prostředí statistického programu R. Je to software určený k manipulaci s daty, výpočtům a grafickému zobrazení dat. Je ideální pro práci s vektory a maticemi, podporuje například příkazy pro výpočet inverzní nebo varianční matice a nabízí přehledné indexování prvků matic. Tento program je založen na rozvinutém programovacím jazyku S. Sám software je také napsán v tomto jazyce, což umožňuje jeho snadnější pochopení a používání. Program je distribuován pod GNU licencí, což znamená že je zdarma ke stáhnutí, dále je velmi snadno rozšiřitelný pomocí balíčků(knihoven), které si můžeme taktéž stáhnout. Balíčky mohou upravit R pro použití například ve statistické genetice, optimalizaci, ekonomii a tak podobně. Dále je možné volat R z Fortranu, C nebo z C++. [6]

Není nutné pro výpočet psát skripty, a tudíž mít větší povědomí o programování, stačí napsat požadovaný příkaz přímo do konzole R. Tento software je multiplatformní (MS Windows, Linux, MacOS), což vylučuje případné problémy uživatelů alternativních operačních systémů.

Příkazy

Nebudeme se zabývat kompletní dokumentací všech příkazů pro R, pouze si vysvětlíme takové z nich, které jsme použili při zpracovávání dat. Díky knihovně MASS provede program diskriminační analýzu po načtení souboru se zdrojovou tabulkou dat (nejlépe ve formátu .txt), v této tabulce sloupce představují statistické znaky a řádky objekty.

`setwd()`	Příkaz změní pracovní adresář R na zadaný adresář v argumentu. např. `setwd("C:/projekty/data/")`
`a=b`	Výrazu nalevo je přiřazena hodnota výrazu napravo např. `pi=3.1415`
`scan()`	Funkce k načítání dat ze souboru do vektoru, v parametru je uveden název souboru; např. `data = scan("data.txt")`
`matrix()`	Funkce konvertuje zadaná data do datového typu matice. Parametr *ncol* udává, do kolika sloupců se mají data načíst, *byterow* určuje, zda-li se data budou načítat po řádcích, nebo po sloupcích. např. `matice = matrix(data,ncol=6,byterow=TRUE)`
`c()`	Vytváří z prvků, zadaných v argumentu, vektor; např. `prvocisla=c(1,2,3,5,7,11,13)`
`colnames()`	Nastaví jména sloupců na názvy, předané argumentem; např. pro matici *skola* o třech sloupcích obdržíme `colnames(skola)=c("1.A","1.B","1.C")`
`as.data.frame()`	Funkce konvertuje matici, danou v argumentu, do datového typu *data frame*, který je nutný pro další zpracování pomocí diskriminační analýzy. např. `data = as.data.frame(matice)`

`library()`	Načte knihovnu, uvedenou v argumentu. V našem případě `library(MASS), která obsahuje příkazy` pro diskriminační analýzu.
`qda()`	Tento příkaz vypočítá diskriminační funkci, pomocí které pak dělíme do skupin. Parametr před tečkou určuje výchozí data, parametr za tečkou určuje skupiny, do kterých budeme zařazovat. např. `disc=qda(matice[,5] .,matice[,-5])`
`predict()`	Funkce provede samotnou diskriminaci, zařadí objekt do skupiny. jako první argument zvolíme diskriminační funkci, ze které budeme vycházet, jako druhý argument zadáme objekty, pro které zjišťujeme skupinovou příslušnost. např. `predict(disc,matice[,-6])$class`

5 Aplikace diskriminační analýzy v internetovém marketingu

Diskriminační analýza nachází uplatnění v mnoha oborech lidské činnosti. Tuto metodu lze využít v rozmanitých vědeckých oborech, například v medicíně (viz výše uvedený příklad rozdělení pacientů do skupin podle závažnosti onemocnění), fyzice (určování typu částice podle její trajektorie v homogenním magnetickém poli), či analytické chemii. Velkým přínosem je diskriminační analýza rovněž v kriminalistice, ale v neposlední řadě také v sociologických výzkumech a marketingu. Podle několika demografických údajů je možné například určit, je-li člověk uživatelem nějáké služby nebo spotřebitelem konkrétního typu výrobku. Reklamní agentury poté mohou upravit reklamní kampaň daného výrobku, tak aby co nejvíce ovlivnila cílovou skupinu. Rozhodli jsme se uvést příklad právě tohoto typu uplatnění diskriminační analýzy.
V současné době dochází k rychlému rozvoji tzv. **instant messaging služeb**, tedy počítačových programů pro internetovou komunikaci v reálném čase.

Mezi tyto počítačové aplikace patří například v České republice velmi oblíbené ICQ, dále pak Skype, Jabber, Windows live messenger (dříve pod názvem MSN) a jiné. Tyto programy obvykle umožňují jak textovou komunikaci, tak internetové volání a posílání souborů. Na platformách mnohých z nich vznikají i různé jednoduché herní aplikace. Mnoho lidí věnuje využívání těchto aplikací značnou část svého volného času. Položili jsme si ale otázku jak silně ovlivňují některé faktory to, zda člověk těchto služeb využívá, či nikoli.
Vybrali jsme pět podle nás nejdůležitějších faktorů. Prvním z nich je **pohlaví**, obecně známým faktem je, že internet využívají více muži než ženy, ovšem to se v poslední době mění a je zajímavé sledovat jak výrazně se to projeví v užívání instant messaging (i.m.) aplikací. Jako druhý směrodatný faktor jsme vybrali **věk**, který je asi nejdůležitějším hlediskem, mladší generace by bez Internetu nebyla sama sebou, a proto je pravděpodobné, že drtivá většina teenagerů bude i.m. služeb využívat, ale můžeme být překvapeni. Třetím statistickým znakem je celkový **počet let vzdělání** (základní a střední škola, popř. vysoké nebo vyšší odborné školy). Vzdělanost člověka má vliv na jeho zájem o nové technologie, mj. tedy i na jeho počítačovou gramotnost, je proto esenciálním faktorem tohoto výzkumu. Dále je možnost připojení k internetu a také míra socializace člověka ovlivněna **velikostí sídla**, ve kterém žije.
U mladší generace se již rozdíly mezi městským a venkovským obyvatelstvem stírají, ale u starších lidí najdeme poměrně znatelné rozdíly ve způsobu života. Posledním, velmi konkrétním faktorem, je **počet hodin týdně strávených na internetu**. U lidí, kteří internet denně využívají k práci či k zábavě, je velká pravděpodobnost, že internet začnou využívat i ke komunikaci s přáteli prostřednictvím zmíněných programů.

Do dotazníku jsme zahrnuli i přímou otázku, zda lidé užívají výše zmíněných služeb, abychom mohli vytvořit diskriminační funkci a určit účinnost diskriminační analýzy. Nečíselným faktorům jsme přidělili logické hodnoty 0 a 1 a velikost sídla jsme převedli na tisíce obyvatel.
Ke sběru dat jsme použili metodu náhodného výběru, ale snažili jsme se zároveň postihnout reprezentativní vzorek populace. Protože je ochota při

vylňování dotazníků obvykle malá, podrobili jsme průzkumu především naše příbuzné a přátele. Například jsme nechali kolovat dotazník po naší třídě, kde nikdo neměl problém odpovědět nám na výše uvedené dotazy. Všichni byli velice ochotní, ale bohužel, pokud by byl vzorek složen pouze ze studentů gymnázia, nebyl by dostatečně reprezentativní. Potřebovali jsme různorodou skupinu lidí, ale kde takovou sehnat. A pak přišel nápad - na fotbale. Fotbaloví fanoušci jsou vskutku malí i velcí a i v ostatních parametrech se mohou navzájem lišit, tak se Petr, mimochodem také velký fanoušek, vydal na utkání klubů Sigma-Kladno. Fandové byli ochotní a získali jsme tak různorodý vzorek populace.
Kompletní data zde bohužel vzhledem k rozsahu práce neuvádíme (kompletní tabulku dat předvedeme na prezentaci), zde je jen ukázka:

Pohlaví	Věk	Počet let vzdělání	Internet	Velikost sídla	Používám IM
muž	19	12	30	100	ano
muž	18	12	10	4.5	ano
muž	37	13	3	1.5	ano
žena	17	12	30	0.3	ano
žena	18	12	30	0.8	ano
muž	18	12	30	5.0	ano
muž	23	18	35	1000	ano
žena	18	12	25	100	ano
žena	50	13	4	100	ne
žena	40	13	2	100	ne
žena	18	12	1	10	ne
muž	84	20	0	30	ne
muž	78	15	1	0.5	ne
žena	74	13	1	0.5	ne
muž	49	19	2	1	ne

Pokud se letmo podíváme na data, zdá se, že jediným dominantním faktorem, ovlivňujícím používání výše zmíněných služeb, je věk, proto jsme jej zkusmo

vyloučili ze zdrojových dat, a ukázalo se, že diskriminační analýza je úspěšná i bez něj. Je to způsobeno velmi vysokou korelací mezi počtem hodin strávených na internetu a používáním i.m. služeb. Potvrdil se tedy náš předpoklad, že lidé, kteří se často baví na internetu, se zde budou bavit i s přáteli. Počet let vzdělání diskriminaci zřejmě výrazně neovlivňuje, pouze u mladších lidí by mohl hrát roli. Role velikosti sídla je také menší, ovšem ukazuje se, že i v dnešní době Internet více využívá urbanizovaná populace.

Nyní si ukážeme jaký je postup v softwaru R pro zařazení nově příchozího objektu respondenta. Jeho charakteristiky zapíšeme do datové tabulky jako první v pořadí, dále spustíme program R a buď ze skriptu, nebo ručně zadáváme příkazy do konzole.

Načteme si knihovnu MASS.

```
library(MASS)
```

Poté nastavíme adresář, ve kterém máme data.

```
setwd("D:/SOC/")
```

Tato data načteme do šesti sloupců.

```
x=matrix(scan("dataDA.txt"),ncol=6,byrow=T)
```

Pro lepší orientaci nastavíme jména jednotlivých sloupců.

```
colnames(x)=c("Pohlavi","Věk","Vzdělání","Hodiny na internetu","Velikost
sidla","Používá i.m.")
```

Data konvertujeme do typu *data frame*

```
x=as.data.frame(x)
```

Vypočítáme diskriminační funkci, klíčová hodnota bude v šestém sloupci, do výpočtu nezahrneme nově příchozího respondenta, který je na první pozici v seznamu.

```
disc=qda(x[-1,6] .,x[,-6])
```

Nakonec zjistíme skupinovou příslušnost funkcí predict.
```
predict(disc,x[1,-6])$class
```

S pomocí těchto příkazů jsme určili typického uživatele instant messaging programů a jeho protiklad.

Užívatel IM	Pohlaví	Věk	Počet let vzdělání	Internet	Velikost sídla
ne	spíše muž	48	14	3	67
ano	muž/žena	22	13	15	44

Pravděpodobnost zařazení nově příchozího jedince do skupiny lidí, kteří nepoužívají IM, je pro tento vzorek dat 0,5125. Naopak pravděpodobnost zařazení mezi uživatele programů bleskové komunikace je 0,4875.
Pravděpodovnost správného zařazení nového objektu do jedné ze skupin byla v tomto případě 0,85. Diskriminační analýza se tedy ukázala vhodnou metodou i pro tuto netradiční aplikaci.
Během práce na této aplikaci nás napadly další možnosti užití diskriminační analýzy na internetu. Například když se správci informační serverů měst či obcí rozhodují, zda umístit na své stránky informace určitého typu (například vztahující se pouze k určité čtvrti), ovšem nechtějí jimi zatěžovat všechny jejich uživatele, mohou dle zájmů, věku či adresy registrovaných uživatelů rozhodnout, zda uživatele tyto informace budou zajímat či nikoli. Takové "výběrové"informace nebo třeba multimediální obsah by pak mohly být poskytnuty jen uživatelům, které tyto informace opravdu zajímají. Takový systém by stránky výrazně zpřehlednil a uživatelů ušetřil spoustu času.
Další z možností, jak využít diskriminační analýzu na internetu, je tzv. age-checking, tedy kontrola věku uživatelů vstupujících na určité stránky. Místo obyčejné otázky typu : "Je vám více než 18 let?", by bylo nutné zodpovědět sérii otázek, podle kterých by byl uživatel zařazen do věkové skupiny. Tento model by mohl být také alternativou k heslům různých účtů nebo pro vstup

na stránky skupin či komunit(například spolužáci.cz). Pokud je při vstupu na stránky položena jediná otázka, často se stane, že ani osoba oprávněná vstoupit na ni nezná odpověď, pokud by otázek bylo více a odpovědi by byly vyhodnoceny pomocí diskriminační analýzy, byl by vstup pro oprávněné téměř jistě zaručen a pro neoprávněné zapovězen.

Závěr

Během naší SOČ jsme rozšířili své obzory v oblasti matematiky, naučili jsme se pracovat s maticemi a pochopili základní teze, metody a pojmy matematické statistiky. Vědy nanejvýš užitečné a pozoruhodné. Seznámili jsme se s velmi perspektivní a účinnou metodou rozdělení objektů do skupin podle jejich vlastností, diskriminační analýzou. Tato metoda nás svou širokou uplatnitelností a relativní jednoduchostí zaujala a určitě se k ní budeme v budoucnu rádi vracet. Jak se ukázalo, diskriminační analýzu lze uplatnit téměř kdekoli. Nemusí sloužit jen v antropologii k rozpoznávání jednotlivých vývojových stádií člověka, ale může nám předpovědět i něco o lidech kolem nás. I když její účinnost byla díky sociologickému zaměření o něco nižší, než jsme původně očekávali, je diskriminační analýza bezesporu metodou, která má v rychle se rozvíjející společnosti se stále větší potřebou rychlého a přesného vyhodnocení velkých statistických souborů zásadní význam. Při psaní této SOČ jsme také pochopili, co to znamená týmová práce. Musíme přiznat, že při naší práci docházelo k drobným sporům, ať už o vědeckou podstatu nebo formální podobu práce. Nakonec byl ale náš „tým“ těmito neshodami utvrzen a všichni jsme začali uvažovat o další práci ve vědeckém kolektivu.

Celkově vzato pro nás práce na této SOČ byla obrovským přínosem. Co si člověk vyzkouší na vlastní kůži, s čím se seznámí blíže než jen z výkladu učitele, pouze o tom se dá říct, že se to skutečně naučil, že to skutečně pochopil. Seznámení s diskriminační analýzou utužilo náš zájem o matematiku a vědu vůbec.

Reference

[1] ANDĚL, J., *Matematická statistika*, Praha: SNTL, 1978

[2] Matematika IV, Popisná statistika [online], Brno: Ústav matematiky FSI VUT, 2005-.[Cit. 15.2.2009]
dostupné z URL: http://mathonline.fme.vutbr.cz/Popisna-statistika/sc-1146-sr-1-a-139/default.aspx

[3] POSPÍŠILOVÁ, L., *Matice*, Brno: MZLU, 2008

[4] SOUKUP. Lekce 7 k DA [online], Praha: FSV Karlovy univerzity, 2007.-[Cit. 23.2.2009]
dostupné z URL: http://samba.fsv.cuni.cz/ soukup/sb037/prednasky/

[5] ŠIK, F., *Lineární algebra zaměřená na numerickou analýzu*, PF MU, 1998

[6] VENABLES,W. N., SMITH, D. M. and the R Development Core Team, *An Introtroduction to R* [online], 2008.-[Cit. 15.3.2009]
Dostupné z URL: http://cran.r-project.org/doc/manuals/R-intro.pdf

Abstrakt

Cílem práce bylo popsat statistickou metodu, zvanou diskriminační analýza, její aplikace v praxi za použití softwaru R, který může urychlit jinak zdlouhavé výpočty.

V první části práce jsme popsali veškeré matematické metody, nutné k použití diskriminační analýzy, což zahrnuje popisnou statistiku a maticovou algebru (operace s maticemi, transpozice, inverze atp.), včetně a její statistické interpretace.

Dále jsme popsali diskriminační analýzu jako metodu, používanou pro klasifikaci objektů s neznámou skupinovou příslušností na základě hodnot příslušných statistických znaků. V práci je jak matematická definice a popis, tak příklady aplikací v širokém spektru oborů. Pro pohodlnější výpočty práce obsahuje též popis statistického softwaru R, včetně nezbytných příkazů.

Nakonec jsme provedli vlastní průzkum na téma instant messanging (tzn. programy bleskové komunikace), ve kterém jsme se snažili najít souvislost mezi věkem, dosaženým vzděláním, počtem hodin strávených u internetu, velikostí sídla, ve kterém respondent žije, a používáním programů bleskové komunikace, jako jsou například ICQ nebo Skype.

Klíčová slova

Maticová algebra, popisná statistika, diskriminační analýza, statistický software R, sociologická studie.