

METODICKÉ POSTUPY ZPRACOVÁNÍ STATISTICKÝCH INFORMACÍ V OBLASTI AGRÁRNÍHO SEKTORU

B. Kába, L. Svatošová

Katedra statistiky, PEF, VŠZ

Anotace :

Příspěvek pojednává o některých robustních deskriptivních charakteristikách, vhodných pro statistickou analýzu populací, jejichž rozdělení není normální. V další části jsou zmíněny moderní grafické prostředky prezentace statistických dat a nejdůležitější transformace, umožňující dosažení přibližné normality rozdělení respektive homogenizace rozptylů zpracovávaných statistických souborů.

Summary :

Several robust statistics based on censored or winsorised samples and techniques of data transformations are presented. This statistical methodology is very advantageous in cases of non-normally distributed samples. Paper outlines also some recent graphical methods for describing statistical data.

Klíčová slova :

Robustní postupy, useknuté charakteristiky , winsorisace, transformace dat.

Key words :

Robust procedures, censored estimates, winsorisation, transformations of data.

Objektivní informace, z nichž mnohé mají charakter statistických dat, jsou v tržní ekonomice nezbytným předpokladem racionálního chování podnikatelů. Je tedy zřejmé, že velmi vzrůstají nároky na kvalitu a věrohodnost disponibilních statistických dat a při jejich zpracování musí být použita adekvátní metodika. Analýzy statistických dat jsou dnes běžně uskutečňovány na samočinných počítačích užitím neustále doplňovaného aplikačního statistického softwaru.

Použití klasických statistických postupů (ať deskriptivních či induktivních), je vázáno na splnění určitých předpokladů. Jsou jimi zejména normalita rozdělení a homogenita rozptylů zpracovávaných datových souborů. I když tyto předpoklady jsou velmi obecné, může v mnoha důležitých případech dojít k jejich značnému narušení. Jedním z možných řešení těchto problémů je užití t.zv. robustních statistických postupů, které dovolují zpracování populací, jejichž rozdělení není normální.

Robustní charakteristiky polohy a variability

Všeobecně užívanými standardními popisnými charakteristikami statistických souborů jsou aritmetický průměr a směrodatná odchylka respektive rozptyl. Pro normální rozdělení jsou tyto charakteristiky přesné a představují vhodný prostředek pro základní orientaci ve zpracovávaných

datech. Pokud však analyzované soubory obsahují odlehlá pozorování respektive jejich rozložení je výrazně asymetrické, jsou výše zmíněné charakteristiky velmi zavádějícími prostředky popisu dat. Jednoduchým robustním protějškem aritmetického průměru je t.zv. *đ-useknutý průměr*. Stanoví se tak, že ze souboru vyřadíme 100đ % nejmenších a největších hodnot a ze zbytku se v zásadě vypočte obvyklý průměr.

Jiným robustním ukazatelem polohy je winsorisovaný průměr. Při jeho výpočtu se žádná hodnota nevyřazuje, ale nejmenší respektive největší hodnoty se nahrazují hodnotami sousedními. T.zv. *g-winsorisovaný výběr* vznikne, nahradíme-li *g* nejmenších pozorování hodnotou stojící na (*g*+1) - místě uspořádaného souboru a *g* největších pozorování hodnotou, jež stojí na (*n-g*) - místě uspořádaného výběru. Některé statistické pakety (např. BMDP) umožňují výpočet winsorisovaného průměru (včetně winsorisovaného intervalu spolehlivosti pro průměr) až do 5-winsorisace.

Jednoduchou robustní alternativou směrodatné odchylky, nezávislou na případných odlehlých hodnotách, je kvartilová odchylka

$$s_k = (K_H - K_D)/2,$$

kde K_H respektive K_D je horní respektive dolní kvartil.

Transformace dat

Validita výsledků indukativní statistické analýzy, uskutečněné pro data, jež nesplňují předpoklady, zmíněné v úvodu předcházejícího odstavce, je velmi diskutabilní. V takových případech se doporučuje provést určitou transformaci dat, jejímž cílem je dosažení alespoň přibližné normality rozdělení a homogenizace rozptylů zpracovávaných souborů. Jistou nevýhodou techniky transformace je obtížná interpretace nově transformovaných dat a jejich jednotek. Pro data týkající se zemědělských respektive ekonomických ukazatelů lze doporučit zejména tyto transformace:

- odmocninová transformace $y = \text{SQRT}(x)$ nebo $\text{SQRT}(x + 3/8)$, jež je vhodná pro data mající charakter absolutních četností vzácně se vyskytujících jevů (tedy pro výběry z Poissonova rozdělení). Pokud malé absolutní četnosti obsahují nuly, osvědčuje se odmocninová transformace ve tvaru $y = \text{SQRT}(x + 0.4)$
- úhlová transformace $y = \arcsin \text{SQRT}(x/n)$ nebo $y = \arcsin \text{SQRT}((x+3/8)/(n+3/4))$, doporučovaná pro data, jež mají charakter relativních četností (procent)
- logaritmická transformace $y = \log x$ případně $y = \log(x+1)$, mají-li data charakter četností

- reciproká transformace $y = 1/x$, užívaná zejména pro proměnné, jež jsou funkcemi času.

Numerické výsledky, uvedené v následující tabulce dokumentují účinnost zmíněných transformací z hlediska homogenizace variability datových souborů.

Transformace dat - příklad

Extremní hodnoty	Variační rozpětí			
	Původní hodnoty	Transformace		
		Odmocninná	Logaritmická	Reciproká
5 - 25	20.0	3.000	0.699	0.160
0.1 - 0.9	0.80	0.632	0.954	8.889
1.2 - 1.6	0.40	0.470	0.125	0.208
56 - 84	28.0	1.682	0.176	0.006

Jednoduché grafické metody prezentace dat

Většina statistických programových paketů poskytuje širokou nabídku grafických vyjadřovacích prostředků. V deskriptivní případně explorační statistické analýze, jejímž cílem bývá orientace ve zpracovávaných datech a formulace určitých pracovních hypotéz, se v poslední době doporučuje prezentace dat ve formě t.zv. lodyhy s listy (Stem and Leaf Display), která shrnuje veškerou informaci obsaženou v datech a umožňuje data kdykoliv rekonstruovat i zpětně analyzovat. Tvoří se tak, že důležité řády zpracovávaných dat se zapisují pod sebe a vytvářejí "lodyhu", nižší řády, jež tvoří "listy", se zapisují do příslušného řádku. Na základě lodyhy s listy lze rychle opticky posoudit kvalitu dat (odlehlá či chybná pozorování), zhodnotit tvar rozdělení (unimodální, polymodální, asymetrie) a rozhodnout se pro případné transformace dat.

Jiným vhodným grafickým prostředkem reprezentace dat je t.zv. krabicový respektive obdélníkový graf (Box and Whisker Plot), který je založen na kvartilech analyzovaného datového souboru a poskytuje velmi názorný přehled o středu a rozptýlenosti rozdělení.

Grafickou analýzu závislostí ve vícerozměrných výběrech umožňují zejména hvězdicové grafy (Star Plot) a paprskové grafy (Sun Ray Plot).

Závěr

Adekvátní reprezentace dat respektive jejich vhodná redukce je předpokladem úspěšnosti statistické analýzy. V situacích, kdy máme pochyby o kvalitě zpracovávaných dat respektive symetrii jejich rozdělení, by měly být více využívány robustní deskriptivní charakteristiky, případně účelné transformace dat. Vhodným doplňkem vstupní numerické statistické analýzy jsou moderní grafické vyjadřovací prostředky. V příspěvku zmíněné metody vizualizace dat umožňují rychlou orientaci v datech, detekci případných odlehlých pozorování i volbu navazujících (zejména induktivních) statistických postupů.

Literatura :

Mc Clave J.T.,Benson P.G. : Statistics for Business and Economics,San Francisco/London,
Dellen/ Macmillan, 1988

Dowdy S.,Wearden S.: Statistics for Research , John Wiley, Inc.,New York , 1983

Koschin F. : STATGRAPHICS , Grada a.s.,Praha 1992

Sachs L. : Applied Statistics, Springer-Verlag, New York,
Berlin, Heidelberg, Tokyo,1978