VYUŽITÍ SYSTÉMŮ STATGRAPHICS PLUS, SAS A MS EXCEL PŘI ANALÝZE ROZPTYLU

TOMÁŠ LÖSTER

Vysoká škola ekonomická v Praze, Fakulta informatiky a statistiky, nám. W. Churchilla 4, Praha 3, 130 67

e-mail: losterto@vse.cz

ÚVOD

Analýza rozptylu je jedna ze základních statistických metod, která zkoumá závislost proměnných. Při výpočtech charakteristik potřebných k analýze rozptylu a následnému testu o závislosti proměnných je možné využít některý z dostupných softwarových produktů. Systémy STATGRAPHICS Plus, SAS a MS Excel umožňují získat charakteristiky potřebné k analýze rozptylu a testu o závislosti proměnných. Je samozřejmé, že výstupy jednotlivých systémů se liší a jsou různě podrobné. K přijmutí závěru testu o závislosti proměnných je nezbytné znát kritickou hodnotu (kvantil F rozdělení), kterou je také možné nalézt v příslušném softwarovém produktu.

ZÁKLADNÍ POJMY

Analýza rozptylu jako základní statistická metoda zkoumání závislosti proměnných může být členěna z různých hledisek. Může být členěna podle počtu faktorů, podle počtu vysvětlovaných proměnných, podle hodnot faktorové proměnné atd.

Jednofaktorová analýza rozptylu zkoumá, zda číselná proměnná Y závisí na slovní (příp. číselné) proměnné X (faktor).

Vícefaktorová analýza rozptylu předpokládá existenci alespoň dvou nebo více třídících faktorů X_i.

Při jednorozměrné analýze rozptylu se předpokládá pouze jedna vysvětlovaná proměnná Y.

Vícerozměrná analýza rozptylu předpokládá v modelu alespoň dvě vysvětlované proměnné.

ANALÝZA ROZPTYLU V JEDNOTLIVÝCH SYSTÉMECH

STATGRAPHICS PLUS - umožňuje velice jednoduchým způsobem aplikovat analýzu rozptylu. Princip je založen na výběru z předem nadefinovaných menu nabídek. Z výstupů je možné zjistit příslušné skupinové průměry, rozptyly, směrodatné odchylky, minima, maxima hodnot číselné proměnné Y atd. V případě jednofaktorové analýzy rozptylu je z výstupu také patrný rozklad celkové variability na meziskupinovou a vnitroskupinovou část, příslušné stupně volnosti, průměrný čtverec, hodnotu testového kritéria a p-hodnota. Průměrný čtverec je označení pro část variability (meziskupinová nebo vnitroskupinová) dělenou příslušným počtem stupňů volnosti. P-hodnota je minimální hladina významnosti, na které je možné zamítnout nulovou hypotézu o nezávislosti proměnné Y na X. Systém STATGRAPHICS Plus také umožňuje pro případ jednofaktorové analýzy rozptylu ověření předpokladů užití této metody. Ve výstupu je možné najít hodnotu testového kritéria Bartletova testu ověřující shodu rozptylů v „k" skupinách a příslušnou p-hodnotu, na jejímž základě lze podle jednoduchého pravidla přijmout závěr o shodě rozptylů v „k" skupinách. V případě, že je p-hodnota > zvolená hladina významnosti (obvykle $\alpha = 0{,}05$), není možné zamítnout nulovou hypotézu Bartletova testu o shodně skupinových rozptylů. Systém také nabízí jako další možnost ověření předpokladů neparametrický Kruskal-Wallisův test. Systém však v případě jednofaktorové analýzy rozptylu neumožňuje ve výstupu nalézt hodnotu poměru determinace. Vzhledem k jednoduchosti výpočtu podle jednoduchého vzorce (podíl meziskupinové variability na celkové variabilitě) je možné tuto hodnotu snadným způsobem dopočítat.

MS EXCEL - umožňuje v jedné ze svých nabídek – analýza dat – jednoduchým způsobem aplikovat analýzu rozptylu. Působení faktorů (proměnných X_i) je chápáno příslušností hodnot číselné proměnné Y buď řádkům nebo sloupcům. MS Excel umožňuje aplikovat jednofaktorovou a dvoufaktorovou analýzu rozptylu. U jednofaktorové analýzy rozptylu je nutné vyznačit vstupní oblast obsahující hodnoty číselné proměnné Y a také vyjádřit, zda faktor je dán jednotlivými řádky nebo sloupci v dané vstupní tabulce. Ve výstupu je vidět, stejně jako u systému STATGRAPHICS Plus, jaké jsou skupinové průměry, rozptyly, počty hodnot v jednotlivých skupinách (n_i). Z výstupu je také zřejmý průběh testu o závislosti proměnných. Kromě stejných údajů jako v systému STATGRAPHICS Plus, tj. rozklad na meziskupinovou, vnitroskupinovou variabilitu, průměrný čtverec, stupně volnosti, hodnotu testového kritéria a p-hodnotu je možné ve výstupu systému MS Excel najít hodnotu kritickou pro daný test o nezávislosti proměnných, tj. kvantil rozdělení F s (k-1) a (n-k) stupni volnosti. Jako je tomu obvyklé u systému MS Excel není ani zde vhodně zvolen překlad jednotlivých termínů. Rozklad variability na jednotlivé složky je označen: „Mezi výběry" – meziskupinová variabilita, „Všechny výběry" – vnitroskupinová variabilita a „Celkem" – celková variabilita. „Rozdíl" je nesprávné označení pro stupně volnosti příslušné části variability, „MS" je zkratka anglických termínů pro průměrný čtverec. Narozdíl od systému STATGRAPHICS Plus není možné ve výstupu najít hodnotu testového kritéria ani p-hodnotu pro Bartletův test, který ověřuje předpoklad užití analýzy rozptylu. Lze však předpokládat, že v případě stejných rozsahů skupin není test o závislosti Y na X příliš citlivý na porušení předpokladů o rovnosti skupinových rozptylů, proto absence průběhu Bartletova testu není zásadním nedostatkem tohoto systému. Stejně jako u předchozího systému není možné ve výstupu najít hodnotu poměru determinace, který určuje sílu závislosti Y na X, a proto je nutné tento poměr speciálně dopočítat. MS Excel nabízí kromě jednofaktorové analýzy rozptylu také dvoufaktorovou analýzu rozptylu (s opakováním nebo bez opakování faktorů).

Dvoufaktorová analýza rozptylu bez opakování předpokládá existenci dvou třídících faktorů Xi s tím, že každá hodnota faktoru je zastoupena pouze jednou.

Tab. č. 1: Příklad vkládání hodnot dvoufaktorového modelu bez opakování

faktor řádkový/faktor sloupcový	hodnota č. 1	hodnota č. 2	hodnota č. 3
hodnota č. 1	hodnoty číselné proměnné Y		
hodnota č. 2			
hodnota č. 3			

Dvoufaktorová analýza rozptylu s opakováním předpokládá existenci dvou třídících faktorů Xi s tím, že každá hodnota faktoru může nabýt několika opakujících se hodnot. Jeden faktor je chápán v řádcích, druhý faktor ve sloupcích.

Tab. č. 2: Příklad vkládání hodnot dvoufaktorového modelu s opakováním

faktor řádkový/faktor sloupcový	hodnota č. 1	hodnota č. 2	hodnota č. 3
hodnota č. 1 (skupina č.1)	hodnoty číselné proměnné Y		
hodnota č. 2 (skupina č.1)			
hodnota č. 3 (skupina č.1)			
hodnota č. 1 (skupina č.2)			
hodnota č. 2 (skupina č.2)			
hodnota č. 3 (skupina č.2)			

Více než dva faktory při analýze rozptylu není možné v systému MS Excel předpokládat.

SAS - ze zmiňovaných programových produktů nejdetailněji umožňuje zkoumat analýzu rozptylu systém SAS. Analýzu rozptylu je možné v tomto systému řešit přes procedury:

- ANOVA (pro vyvážené modely)
- KLM (pro vyvážené a nevyvážené modely)
- MIXED (pro smíšené modely)
- Neparametrická ANOVA

Vyváženým modelem je chápán takový model, kdy ve skupinách vzniklých podle působení příslušných faktorů je stejný počet statistických jednotek. Neparametrická ANOVA je vhodná zejména pro malé soubory, u kterých se nedá ověřit podmínka o normalitě rozdělení vysvětlované (závislé proměnné) v jednotlivých skupinách.

V případě vyvážených modelů lze použít proceduru ANOVA. Princip vkládání vstupných hodnot je podobný systému STATGRAPHICS Plus. Pro případ jednofaktorové analýzy rozptylu výstup obsahuje základní informace o souboru hodnot, stejně jako v případě systémů STATGRAPHICS Plus i MS Excel. Mj. lze z výstupu získat průměrné hodnoty v rámci skupin, směrodatné odchylky, průměrné čtverce (chápané stejně jako v systémech STATGRAPHICS Plus i MS Excel) atd. Samozřejmě je patrný rozklad na meziskupinovou i vnitroskupinovou variabilitu a průběh testu o závislosti proměnné Y na Xi. Na rozdíl od předchozích systémů je zde uvedena hodnota poměru determinace, která je označena jako R-square. Ve výstupu je také uveden Bartletův test zkoumající shodnost rozptylů v k skupinách a průběh Levenova testu a Brownova-Forsythova testu zabývající se rozptylem.

Pro případ, že je model nevyvážený, je možné dokázat, že je vhodné použít proceduru označenou KLM, neboť výsledky F statistiky mohou u procedury ANOVA vycházet rozdílně.

ZÁVĚR

Z uvedených postupů je patrné, že jednotlivé systémy se liší nejen podrobností výstupů, způsobem vkládání hodnot ale také možností aplikace dané procedury. Nejmenší šíři aplikace má systém MS Excel, který umožňuje aplikovat analýzu rozptylu do maximálního počtu dvou faktorů, naopak nejvíce propracován je z hlediska analýzy rozptylu systém SAS. Systém STATGRAPHICS Plus umožňuje aplikovat analýzu rozptylu pro případ, že se nejedná o příliš složité modely. Jedná-li se o složité modely, pak zejména u síťového provedení systém STATGRAPHICS Plus může selhávat. Je zřejmé, že analýza rozptylu bez použití softwarových produktů je značně pracná, někdy i nemožná. Z tohoto důvodu je vhodné využívat některý z nabízených systémů. Je však nutné nejen mechanicky aplikovat určité postupy pomocí menu nabídek ale také je třeba prostudovat přiloženou dokumentaci a ověřovat příslušné předpoklady užití testů i v případě, že je daný softwarový produkt nenabízí.

LITERATURA

[1] ARLTOVÁ, M., BÍLKOVÁ, D., JAROŠOVÁ, E., POUROVÁ, Z.: Příklady k předmětu statistika A, VŠE, Praha 2003.

[2] CHAJDIAK, J.: Štatistické úlohy a ich riešenie v Exceli, Statis, Bratislava 2005.

[3] JAROŠOVÁ, E., PECÁKOVÁ, I.: Příklady k předmětu statistika B, VŠE, Praha 2000.

[4] MAREK, L., a kol.: Statistika pro ekonomy aplikace, Profesional Publishing, Praha 2005.

[5] ŘEZANKOVÁ, H.: Analýza kategoriálních dat, VŠE, Praha 2005.