Jan Řehák, Ondřej Brom

SPSS

Praktická analýza dat

Od základních po pokročilé statistické techniky

Příprava dat a úprava výstupů

Vhodné do škol s výukou SPSS

Pro statistiky, analytiky, sociology i manažery

computer press

Jan Řehák, Ondřej Brom

SPSS – Praktická analýza dat

Computer Press
Brno
2015

SPSS – Praktická analýza dat

Jan Řehák, Ondřej Brom

Obálka: Martin Sodomka
Odpovědný redaktor: Roman Bureš
Technický redaktor: Jiří Matoušek

Objednávky knih:
http://knihy.cpress.cz
www.albatrosmedia.cz
eshop@albatrosmedia.cz
bezplatná linka 800 555 513
ISBN 978-80-251-4609-5

Vydalo nakladatelství Computer Press v Brně roku 2015 ve společnosti Albatros Media a. s.
se sídlem Na Pankráci 30, Praha 4. Číslo publikace 23 277.

1. vydání

ALBATROS MEDIA a.s.

Obsah

Pracovní soubory ke stažení 11

Předmluva 13

Úvod 15

O programu 19

ČÁST I

PŘÍPRAVA DAT

Před analýzou dat 30

KAPITOLA 1

Soubory 31

Manuální zápis dat do souboru 31
Převzetí datového souboru do programu 35
Vybavení souboru – Variable View 36
Datasety 40
Transpozice 41
Restrukturace 43
Spojování souborů 52
Agregace případů 56

KAPITOLA 2

Případy 61

Manuální úpravy 61
Uspořádání případů 62
Výběr případů – práce s podmnožinou záznamů 63
Štěpení souboru pro přímou práci 67
Vážení 68

KAPITOLA 3
Proměnné **71**

Transform **71**
Změna existující a tvorba nové proměnné výpočtem **73**
Rekódování **75**
Počet výskytů **78**
Pořadí **80**
Třídní intervaly **82**
Automatické rekódování **85**
Konstrukce dummy proměnných **86**
z-skóry **88**

ČÁST II
STATISTICKÉ TABELACE A ANALÝZY

Od jednoduchého přehledu k vícerozměrné analýze **90**

KAPITOLA 4
Statistické tabelace a přehledy **91**

Analyze – … **91**
Codebook – rychlý přehled vlastností jednotlivých proměnných **92**
Case Summaries – výpisy a sumarizace dat **95**
Frequencies – tabulky četností pro kategorizované proměnné **97**
Descriptives – základní popisné statistiky **99**
Means – tabulky statistik ve skupinách **101**
Explore – popis rozložení pomocí kvantilů **105**
Ratio – výpočet a testování poměrových statistik **110**
Multiple Response **113**

KAPITOLA 5
Testování komparačních hypotéz **119**

Analyze – … **119**
Crosstabs – kontingenční tabulky: komparace četnostních distribucí a asociace nominálních a ordinálních proměnných **120**

One-Sample T test – testování průměru s vnějším kritériem **127**
Independent-Samples T test – porovnání průměrů dvou souborů **128**
Paired-Samples T test – porovnání průměrů u dvou proměnných jednoho souboru **131**
One-Way ANOVA – komparace průměrů více souborů **133**
Neparametrické testy – analýza založená na pořadí **139**
A) Nonparametric Tests: One Sample **140**
B) Nonparametric Tests: Independent Samples **148**
C) Nonparametric Tests: Related Samples **153**
Nonparametric Tests: Legacy Dialogs **156**
A) Procedura Legacy Dialogs – Chi-square – test dobré shody chí-kvadrát **158**
B) Procedura Legacy Dialogs – Binomial **158**
C) Procedura Legacy Dialogs – Runs **159**
D) Procedura Legacy Dialogs – 1-Sample K-S **160**
E) Procedura Legacy Dialogs – 2 Independent Samples **161**
F) Procedura Legacy Dialogs – K Independent Samples **162**
G) Procedura Legacy Dialogs – 2 Related Samples **162**
H) Procedura Legacy Dialogs – K Related Samples **164**

KAPITOLA 6

Vícerozměrná statistická analýza **165**

Analyze – … **165**
Korelační analýza – procedura Bivariate **166**
Lineární regresní analýza – procedura Linear **168**
Vyhlazení dat křivkou – procedura Curve Estimation **173**
Optimální redukce vícerozměrné informace a hledání vnitřních příčin variability datového vektoru – procedura Factor **179**
Seskupování objektů podle podobností jejich profilů – procedura Hierarchical Cluster **183**
Seskupování objektů podle podobností jejich profilů – procedura K-means Cluster **187**
Vlivy vnějších faktorů na variabilitu číselné proměnné – procedura Univariate **193**

ČÁST III

VÝSTUPY A JEJICH ÚPRAVY

Editace výstupu a efektivní předání výsledků uživatelům analýzy 202

KAPITOLA 7

Výstupní okno – Viewer 203

Struktura výstupního okna 203
Objekty výstupního okna 205
Otevření a používání výstupního okna a směrování objektů do výstupních oken 206
Úpravy a organizace výstupního okna 206
Hromadná úprava objektů výstupního okna 208
Podmíněné formátování (Conditional Styling) 210
Kopírování objektů okna do externích aplikací 212
Export celého výstupu nebo jednotlivých objektů 213
Nastavení výstupního okna 214
Výstupní okno v aplikaci Smartreader 214

KAPITOLA 8

Pivotní tabulky 217

Struktura pivotní tabulky 218
Oblasti pivotní tabulky 218
Editace pivotní tabulky 219
Označení polí pro editaci 220
Změna struktury pivotní tabulky – pivotace 220
Změna pozice řádků a sloupců 221
Odstranění sloupců a řádků nebo jejich skrytí 222
Vytváření nových sloupců a řádků 222
Seskupování řádků nebo sloupců 223
Seřazení řádků 223
Změna šířky sloupců 224
Úprava obsahu a vzhledu jednotlivých polí 224

Úprava vlastností tabulky 225
Šablona tabulek 226
Doplnění nadpisu tabulky, komentáře a poznámky pod čarou 227
Vytvoření grafu z tabulky 228
Výchozí nastavení tabulek 229

KAPITOLA 9
Grafická vizualizace dat 231

Grafy v IBM SPSS Statistics 232
Typy a zadávání prezentačních grafů 233
Obecné volby při tvorbě grafů 233
Sloupcový graf (Bar) 235
3-D sloupcový graf (3-D Bar) 238
Spojnicový graf (Line) 239
Plošný graf (Area) 240
Kruhový (koláčový) graf (Pie) 240
Graf rozpětí (High-Low) 240
Graf rozptýlení – krabicový graf (Boxplot) 242
Graf rozptýlení – intervalový graf (Error Bar) 243
Populační pyramida (Population Pyramid) 243
Bodový graf a bodový graf hustoty (Scatter/Dot) 244
Histogram (Histogram) 245
Sekvenční graf 245
PP a QQ grafy 246
Paretův graf 246
Grafy kontroly kvality – regulační diagramy (control charts) 247
Editace grafu z prezentační grafiky 247
Editační okno grafu – Chart editor 248
Doplnění objektů do grafu 249
Editace grafu nebo jeho objektů z nabídky 250
Výběr objektů grafu pro editaci 250
Editace objektů grafu v editačním okně a jejich odstranění 251
Editace objektů v okně vlastností 252

Zvláštní módy editačního okna 255
Šablony grafů 255
Volby nastavení grafů pro práci 256
Chart Builder 257
Graphboard Template Chooser 257

APENDIX A
Syntaktický jazyk 261

Struktura syntaxe 262
Jazyk syntaxe 263
Proměnné 265
Klíčová slova mimo dialogová okna 265
Nápověda k syntaxi – struktura příkazu v nápovědě 268
Editor syntaxe 270
Syntaxe ve výstupovém okně a žurnál 272
Efektivní práce se syntaxí 277

APENDIX B
Funkce kalkulačky pro transformace proměnných (Compute Variables, Select Cases) 279

Dialogové okno kalkulačky 279
Pravidla zápisu vzorců v kalkulačce procedury Transform – Compute Variables 281
Transformační postupy v syntaktickém jazyce 282
Přehled funkcí a konstant systému 286
Arithmetic functions – aritmetické funkce 286
CDF & Noncentral CDF – kumulativní distribuční funkce 287
Conversion – konverze formátů 288
Current data and time – aktuální datum a čas 288
Date Arithmetic – operace s daty 289
Date Creation – tvorba proměnných data 289

Date Extraction – extrakce data 289
Inverse DF – inverzní distribuční funkce 290
Miscellaneous – různé funkce 290
Missing Values – funkce chybějících hodnot 290
PDF & Noncentral PDF – hustoty pravděpodobností a pravděpodobnostní funkce 291
Random Numbers – generování náhodných čísel 291
Search – vyhledávací funkce 291
Signifikance – výpočet dosažené statistické významnosti 292
Statistical – statistické funkce pro data v řádku (vybrané proměnné) 292
Scoring – skórovací formule 293
String – funkce textových proměnných 293
Time Duration Creation – tvorba proměnných délky časového intervalu 295
Time Duration Extraction – extrakce proměnných délky časového intervalu 295

APENDIX C
Přehled modulů IBM SPSS Statistics 297

Obsah a role modulů systému 297
Analytické doplňky 298
Sdílení výstupů 298

APENDIX D
Přehled procedur IBM SPSS Statistics Base 299

Procedury záložky *Data* v IBM SPSS Statistics Base 299
Procedury záložky *Transform* v IBM SPSS Statistics Base 301
Procedury záložky Analyze v IBM SPSS Statistics Base 301

APENDIX E
Přehled procedur v jazyce Python zařazených do IBM SPSS Statistics 305

APENDIX F
Přehled procedur v jazyce R zařazených do IBM SPSS Statistics 309

Literatura externí 313

Manuály IBM SPSS 313
Acrea CR Výukové materiály 314

Rejstřík 315

Obrazová příloha 327

I – Tlačítka pro práci se systémem část 327
II – Úprava vzhledu pivotních tabulek pomocí šablon 329
III – Sloupcový graf – dvojí uspořádání téže základní informace 330
IV – Třírozměrný sloupcový graf 331
V– Kruhový (koláčový) graf s 3D efektem 331
VI – Hi-Lo graf ve dvou uspořádáních kategorií: a) oficiální seznam krajů, b) pořadí krajů podle klesajícího procenta u ČSSD 332
VII – Dvě varianty souřadnicového grafu: a) graf s proloženým trendem a pojmenovanými odlehlými hodnotami, b) graf s boxploty marginálních statistických řad 333
VIII – Maticový souřadnicový graf s histogramy jednotlivých vstupů 334
IX – Komparace oblastí v krabicovém grafu pro skupinku tří stran 335
X – Kartodiagram 335
XI – Hvězdicový graf 336

Pracovní soubory ke stažení

Soubory použité v knize jsou k dispozici ke stažení na stránkách knihy na adrese ***http://knihy.cpress.cz/K2213*** pod odkazem **Soubory ke stažení** nebo alternativně na stránkách autorů na adrese ***www.acrea.cz/kniha***.

V archivu naleznete soubory:

- ***EHS v ČR.sav*** – část souboru evropského výzkumu hodnot
- ***Kraje 2013 - volby profily.sav*** – krajské volební zisky parlamentních stran ve volbách do PS Parlamentu ČR 2013
- ***Kraje 2013 - volby.sav*** – krajské volební zisky parlamentních stran ve volbách do PS Parlamentu ČR 2013
- ***Měření_hmotnosti.sav*** - soubor s účastníky dietologické stude
- ***Obvody Prahy 2012 - charakteristiky.sav*** – vybrané demografické charakteristiky správních obvodů Prahy z roku 2012
- ***Okresy 2009 2012.sav*** – vybrané demografické údaje z let 2009 a 2012 v okresech a volební zisky parlamentních stran ve volbách do PS Parlamentu ČR 2013
- ***Okresy 2010 - volby.sav*** – okresní zisky parlamentních stran ve volbách do PS Parlamentu ČR 2010
- ***Okresy 2013 - volby.sav*** – okresní zisky parlamentních stran ve volbách do PS Parlamentu ČR 2013
- ***Okresy 2013.sav*** – vybrané demografické údaje z let 2009 a 2012 v okresech a okresní volební zisky parlamentních stran ve volbách do PS Parlamentu ČR 2010 a 2013
- ***Okresy mimo Prahu 2012 - charakteristiky.sav*** – vybrané demografické charakteristiky mimopražských okresů z roku 2012
- ***Podnik.sav*** – soubor s údaji o zaměstnancích fiktivního podniku
- ***Sales.sav*** – soubor z výzkumu spokojenosti s obchodním řetězcem
- ***Transakce.sav*** – transakční soubor s položkami nákupu v obchodním řetězci

Předmluva

Knihy pojednávající o SPSS jsou ve velké většině laděny jako učebnice statistiky, u nichž je výklad statistických metod svázán s aplikacemi softwaru. Poskytují výhodu spojení statistické znalosti s ovládáním spolehlivého prostředku pro analýzu dat, a tudíž plní dva účely současně. Nevýhodou přístupu je však to, že soustředění se na statistické procedury programu nutně zanedbává (ve výuce i v praktické činnosti) jiné potřebné role, které takový prostředek musí mít. Jsou to především dvě fáze analytické práce: příprava dat a manipulace s výstupy.

Při své dlouholeté pedagogické i konzultační činnosti jsem při práci s programem (téměř čtyřicet let) zjišťoval, jak málo si jsou uživatelé i učitelé vědomi jeho bohatých praktických možností při přípravě dat i při úpravě výstupů. Přitom je to jedna z nejpodstatnějších vlastností programu: postupy, které ulehčují a zrychlují (někdy nudnou a nezáživnou a časově náročnou) práci v těchto nutných aktivitách datového zpracování. Proto jsme se rozhodli pro přístup, který dá vystoupit bohatství systému pro všechny aktivity analytika. Rozhodli jsme se pro důraz na to, co se jinde hledá obtížně: komplexní přípravu datového souboru v počáteční etapě i v průběhu a po ukončení analýzy a na funkce, které jsou potřebné v průběhu interakce „uživatel – data – analýza – výstupy".

Pokusili jsme se připravit knížku, která by sloužila pro studenty ve výuce a pedagogickou práci učitelů (kurzy softwaru, praktika ze statistiky, příprava závěrečných prací), jako příruční přehled pro konkrétní práci analytika či vědeckého pracovníka i jako vstup do programu pro nové uživatele. Našimi cíli bylo poskytnout knižní formu podpory uživatelů: a) rychlé seznámení se s jednotlivými procedurami a s možností proklikat se všemi jejich možnostmi, b) příruční/referenční přehled pro průběžnou práci, c) pohled na to, co je velkou předností programu, ale je málo využíváno, d) manuál v českém jazyce.

Velký rozsah systému vedl ovšem k nutné redukci popisovaných procedur. Nejvíce je redukována část statistické procedury, avšak všechny základní a běžné procedury a metody jsou zahrnuty. Vynechali jsme ty metody, které svojí složitostí potřebují již určitou analytickou a výpočetní zkušenost, a proto pro jejich uživatele nebude obtížné tyto procedury (ovládané zcela analogicky jako ty jednodušší) aplikovat. Nemohli jsme také z důvodů prostorových limitů uvést různé, i když nesmírně užitečné obslužné funkce a všechny postupy zajišťující návaznosti a přechody vně programu.

Obsah knihy je založen na verzi 23 systému. Vše, co jsme zahrnuli, však má trvalejší platnost, v následných vyšších verzích může jít o obohacení a rozšíření jednotlivých procedur, současné bohaté funkce však budou zachovány.

Systém IBM SPSS Statistics je nejrozšířenějším a nejpoužívanějším statistickým prostředkem nejen u nás, ale i ve světě. Důvod je v principu jeho vývoje: byl rozvíjen po celou dobu od roku 1968 nejen podle novinek statistické teorie, ale především pro potřeby uživatelů a podle jejich požadavků. Za dobu své existence každý rok přichází s vyšší rozšířenou verzí a dosáhl opravdu velmi širokého rozsahu. Velmi rozsáhlé portfolio možností a jednoduchá uživatelská forma vede

k tomu, že a) nikdo nezná systém do všech detailů, b) každý si najde to, co potřebuje a c) standardní postupy jsou k dispozici velmi snadno a bezproblémově.

Sama statistická věda se rychle rozvíjí a nabízí stále nové metody, praktické aplikace se rozvíjejí a neustále vznikají nové, kvalifikace uživatelů pro analytickou práci se zvyšuje a rozšiřuje. Procesy datových analýz se stávají nutnou podmínkou úspěchu v soudobém informačním světě. Věřím, že touto publikací přispějeme k ulehčení práce pro nové uživatele. Věřím, že přispějeme k pracovnímu komfortu uživatelů i k úplnějšímu využívání všech předností systému a tím i k úspěšným výsledkům.

Praha, červenec 2015

Jan Řehák

Úvod

Co potřebuje analytik v praxi?

U univerzálního statistického programu předpokládáme tři zásadní splněné podmínky:

a) *statistická stránka:* je statisticky korektní, numericky a algoritmicky přesný, poskytuje správné a prověřené metody a obsahuje systém metod pro základní otázky analýzy dat v různých oborech aplikací,
b) *uživatelská stránka:* je uživatelsky příjemný a je koncipován tak, aby usnadňoval praktický proces analýzy v plné šíři interakce uživatele s daty,
c) *vnější kontext vývoje:* neustále se dynamicky rozvíjí podle potřeb doby.

K tomu přistupuje ještě *cena za výkon a obsah podle potřeb uživatele* (tedy nikoliv cena jako taková). **IBM SPSS Statistics** splňuje tyto podmínky už od svého vzniku v roce 1968 a to také bylo vždy důvodem jeho vysoké oblíbenosti.

A. Statistická korektnost je *podmínkou naprosto nutnou*. Výběr metod není jednoduchý, u sofistikovaných postupů záleží nejen na teoretických vlastnostech odvozených matematickou statistikou, ale také na volbě algoritmů a numerických postupů. A je z čeho vybírat – za svoji existenci statistická věda vyvinula tisíce metod a postupů, koeficientů, způsobů prezentace. Ne všechny používáme, některé se neukázaly vhodné, některé nebyly přijaty do hlavního proudu a byly zapomenuty (mnohdy neprávem), některé jen paralelně řešily to, co už bylo dobře zavedeno jinak.

U některých úloh existuje řada přístupů a algoritmizací a situace výběru není snadná. Některé procedury v SPSS byly proto designovány a programovány na specializovaných prominentních akademických pracovištích.

Velmi také záleží na specifických zvyklostech i potřebách jednotlivých oborů. Program SPSS byl vždy vyvíjen v konzistenci s přáními uživatelské komunity. A navíc pod průběžnou systematickou kontrolou uživatelů (jednotlivců i univerzitních kateder), takže každá chyba byla rychle nalezena. Portfolio nabízených postupů vychází tedy nejen z představ teoretiků, ale bylo vždy určováno do velké míry požadavky praxe.

B. Co znamená pojem „*uživatelsky příjemný*"? Především, a tak to bylo v průběhu let vždy chápáno, je to *snadné ovládání*. Už při vzniku nabídl tento program *uživatelsky orientovaný, mnemotechnicky založený syntaktický jazyk zadávání (syntaxe)*, který se osvědčil. Byl jedním z aspektů, který předznamenal úspěch programu u širokého okruhu uživatelů – je proto k dispozici a je rozšiřován dodnes.

Později, s nástupem Windows, bylo rychle zavedeno přehledné a jednoduché *zadávání pomocí oken*. Uživatel si proto může *vybrat*: řízení programu okny nebo syntaxí. To je zcela věcí vkusu a osobní preference.

C. Uživatelská příjemnost („*user friendly*“ program) ale znamená i další momenty, které jsou pro analytika podstatné. Pohodlí analýzy znamená, že máme v jednom analytickém běhu k dispozici vše, co je potřeba. Vše je po ruce a kdykoliv to můžeme použít: zavádění nových proměnných a překódování či transformaci původních, výběry podsouborů a návraty k původnímu souboru či přechod k jiným podsouborům, opakované výpočty na podsouborech, rychlá změna parametrů procedury, spojování souborů, agregace, rychlé přechody mezi soubory, zavádění a rušení vah apod.

Důležité jsou také jednoduché návaznosti procedur, přecházení s výsledky jedné procedury do druhé a využití výsledků pro další analýzu, (velmi podstatné) rychlé opravy omylů při zadání či při vývoji modelů a upřesňování postupu; a také změny ve výstupech a jejich úpravy. Souběžné otevření několika datových souborů a přímé přecházení mezi nimi jen dalším aspektem, který skýtá analytické pohodlí.

Uživatelská příjemnost je tedy forma nabídky, která zrychluje, zjednodušuje postup a pomáhá analytikovi bez potíží a zdržování dojít k výsledku. Nenutí koncentrovat se na techniku zadávání, ale uvolňuje myšlenkovou kapacitu na úlohu, řešení, volbu metod, soustředění na další kroky. Patří sem však též jednoduché napojení na vnější zdroje dat a rychlá publikace výsledků mimo systém.

Dalším aspektem uživatelské příjemnosti systému je otevřenost systému ve všech směrech:

- přebírání (a předávání) různých formátů dat – přímé i cestou ODBC,
- rozšiřování nabídkových menu o okna vlastních výpočetních procedur či výstupových modifikací a doplňků – makra systému, skripty napsané v jazyku Python, procedury v R,
- napojování s přechody do a z jiných uzavřených programů – např. Amos.

Rozsáhlá uživatelská pomoc *Help* popisuje užití jednotlivých voleb v procedurách, algoritmy, výukový text.

Práce s programem **IBM SPSS Statistics** se v mnohém podobá běžné praxi, na kterou jsme zvyklí ze standardních programů pro OS Windows. Ovládá se pomocí menu, oken a ikon. Program je ovšem uzpůsoben speciálnímu úkolu, pro nějž byl vytvořen. Nabídková okna obsahují statistické postupy a jsou optimálně uzpůsobena analytické práci. Doprovodný syntaktický jazyk je jednoduchý a uživatelsky příjemný.

D. Vývoj informačních technologií a rozvoj matematiky a matematické statistiky znamená i tlak na naše statistické programy. Doba mění, rozvíjí a přináší nové požadavky a potřeby, ale také výsledky:

- Rozvoj nových statistických metodologií přináší nové postupy, které zpřesňují modely reálného světa. Teorie statistiky není sprintem, je to pozvolný, ale stálý proud nových vědeckých poznatků, vývoj nových i revize a prohlubování běžných tradičních postupů. Do nativních procedur **IBM SPSS Statistics** jsou zařazovány metody prověřené, otevřenost systému však otevírá možnost připojit jakékoliv procedury z literatury i z vlastního vývoje.
- Stále silnější a rychlejší hardware a s ním spojený software operačních systémů nutí přizpůsobovat se i softwaru aplikačnímu, otevírá ale cesty těm postupům, které byly ještě nedávno neúnosně zdlouhavé – hodiny se postupně zázrakem změnily v minuty, minuty v sekundy.
- Rychle se měnící požadavky aplikačních úloh, potřeby tvůrců i uživatelů informace v jednadvacátém století vedou k potřebě softwarových opatření: vytvořené mohutné masivy stát-

ních i podnikových dat, Big Data, rychlý sběr ad hoc dat, průběžné záznamy dat z procesů. Zrychlená možnost analytických závěrů vede přirozeně k formulaci zcela nových analytických otázek a úloh, k automatizaci analýz, široké aplikaci dávkových i on line rozhodovacích procesů, k rozvoji oboru *Predictive Analytics*, a s tím vším rostoucí vzdělanost současných i potenciálních uživatelů. Nejzásadnějším požadavkem doby je však rychlost zpracování a automatizace – informace zastarává rychle, rozhodování musí probíhat v reálném čase, náklady na čas zpracování je nutno minimalizovat.

Vývoj softwaru **IBM SPSS Statistics** se zaměřuje na to, aby technické aspekty analytické práce co nejméně narušovaly proces statistické aplikace samotné a abychom se mohli věnovat substantivní stránce, výsledkům, korektnímu nasazování technik, vhodnosti výstupů – tedy aby mohly při vytváření závěrů *„méně pracovat prsty a myš a více mozek"*.

Stále složitější modely a algoritmy, umožněné hardwarem, vedou k velkému rozsahu systému, a tudíž i k zvýšené náročnosti na rozvoji údržbu a náklady. Proto k výhodám patří také *„samostatná modularita"*, která znamená, že uživatel si pořídí jen tu část komplexu speciálních modulů, která odpovídá jeho osobním aplikačním potřebám. Modulární systém pracuje jako jeden nedílný celek v té sestavě, kterou si uživatel vybere.

Navíc ale každý modul (kromě modulů, které mají obslužný charakter jiných statistických procedur) může fungovat sám, a to s plným vybavením datových úprav (které byly dříve jen v modulu **Base**) a s plně funkčním výstupovým oknem **Viewer**. Kromě toho je k dispozici **Developer**, který obsahuje všechny vstupní, modifikační a výstupové funkce, ale neobsahuje žádné statistické procedury a slouží těm, kteří potřebují pouze připravovat datové soubory a prezentovat vhodně výsledky. Uživatelé procedur v jazycích Python nebo R či C++ tu mají manipulační datový základ a výstupní editor, do kterého mohou vkládat své vlastní procedury a vytvořit si své vlastní systémy.

V této knize popisujeme modul **Statistics Base**. Věnujeme ale obzvláštní pozornost procedurám přípravy dat (Část 1) a výstupům (Část 3), proto je přehled užitečný i pro samostatné užívání jiných modulů a pro aplikace **Developeru**. Část 3 je také určena pro ty, kdo nezpracovávají data, ale přebírají výsledky analýz volným samostatným (a bezplatným) výstupovým modulem **Smartreader** a chtějí výsledky dále editovat.

Při výběru procedur pro tuto knihu (celý obsah systému není možné rozumně vměstnat do rozumného objemu) jsme vycházeli ze tří předpokladů:

a) Kniha má být příručkou pro praktiky a studenty, kteří nemají specializované IT nebo matematické vzdělání, ale provádějí konkrétní analýzy dat – proto volíme detailní postupy.
b) Podle našich konzultačních a pedagogických zkušeností si uživatelé plně neuvědomují možnosti datových úprav a editace výstupů – proto části 1 a 3 popisujeme co nejúplněji.
c) U statistických procedur se zaměřujeme na běžné a základní metody, které jsou v analýze nejčastěji používány – u složitějších metod je třeba vyšší statistická znalost a jistá zkušenost nebo absolvování tematického kurzu, avšak poté je zadávání zcela mechanické a obdobné nebo jednoduše návodné.

Z témat analýzy jsme byli nuceni vynechat postupy časově-prostorových analýz a predikcí, analýzu spolehlivosti měření, mnohorozměrné škálování, dvoukrokové seskupování, ordinální regresi, proceduru lineárních modelů a některé další. K těmto tématům odkazujeme čtenáře na manuál programu.

Knihu jsme psali pro širokou uživatelskou komunitu systému, který funguje a je oblíben již čtyřicet sedm let a zajišťuje tradici, kvalitu a rozvoj. Využili jsme své i firemní dlouholeté zkušenosti z výuky a analytické práce s programem. Děkujeme svým kolegům ze společnosti ACREA CR – podpořili naši snahu trpělivostí s naší částečnou absencí v běžných odborných činnostech a jejich bohaté lektorské, konzultační, analytické znalosti jsme využili v zásadních i dílčích rozhodnutích.

O programu

Programový systém **IBM SPSS Statistics** je speciální programový systém pro statistické zpracování dat, který zahrnuje techniky a postupy pro práci s úpravami datových souborů, metody statistické analýzy, editační úpravy výstupů a mnoho způsobů, jak zrychlit, zjednodušit a zefektivnit cestu od vstupu dat k závěrečné zprávě či k prezentaci výsledků a k publikaci. Od roku 1968, kdy byla k dispozici jeho prvotní, velmi jednoduchá verze, až do dneška vždy patřil k nejrozšířenějším a nejoblíbenějším. Důvodem k tomu bylo od počátku jeho příjemné uživatelské rozhraní, v té době zcela inovativní. A po celou dobu existence vykazoval systém vždy jednoduché ovládání a uživatelské prostředí.

Program se nejdříve orientoval na sociální vědy, ale už ve verzích na mainframe počítače rychle opustil tuto doménu a stal se univerzálním statistickým systémem pro analýzu dat, používaným ve všech oborech. Pro svoji jednoduchost je oblíben nejen analytiky bez profesionálního statistického vzdělání, ale i pro výuku studentů. Je běžnou výbavou výzkumných firem. K přednostem programu patří to, že skýtá různé způsoby ovládání, a proto si každý uživatel může vybrat ten způsob, který mu vyhovuje.

Modularita systému

Program **IBM SPSS Statistics** je **modulární systém**, jehož základní část Base je jádrem aplikací a obsahuje běžné standardní postupy analýzy dat. Na něj navazují další moduly, které mají speciální charakter – buď analytický, nebo obslužný. Vznikaly historicky, tak jak se vyvíjely potřeby analytické práce a požadavky uživatelů. Návazné moduly jsou zaváděny odděleně proto, že je nepotřebují všichni uživatelé a jejich metody a postupy vyžadují speciální znalost a nasazení v praxi.

Většina modulů může ale fungovat samostatně, je vybavena všemi obslužnými procedurami základu Base a to jak v práci s úpravami dat, tak ve výstupní části **Viewer**. Např. analytik, který potřebuje pouze analýzu a predikci v časových řadách, si může zakoupit jen **IBM SPSS Statistics Forecasting**, ten, kdo má za úkol jen připravovat data pro další analytiky, si může vystačit s modulem **IBM SPSS Data Preparation**.

Tabulka 1 Moduly systému IBM SPSS Statistics

Název modulu	Role v systému
Statistics Base	příprava dat, základní tabelace, statistické metody, grafy
Custom Tables	vytváření komplexních tabulek na obrazovce
Data Preparation	techniky pro přípravu a kontrolu kvality dat
Exact Tests	přesné statistické testy pro neparametrické techniky a tabulky četností
Regression	regresní postupy (mimo lineárního modelu)
Advanced Statistics	pokročilé statistické metody
Categories	metody analýzy korespondencí

Název modulu	Role v systému
Forecasting	analýza a predikce časových řad
Complex Samples	plánování a zpracování pravděpodobnostních výběrů
Conjoint Measurement	plánování a analýza metodou sdružených měření
Decision Trees	metody rozhodovacích a asociačních stromů
Neural Networks	neuronové sítě
Direct Marketing	segmentace, RFM analýza, skórování, plánování kampaní, profilování
Missing Values	analýza a imputace chybějících údajů
Bootstraping	metoda odhadu parametrů nezávislá na normálním rozložení

Každý z modulů obsahuje nativní procedury systému, v menu jsou ale také vloženy vnější procedury programované v jazycích Python nebo R, které nabízejí doplňkové a speciální metody zpracování dat. K systému se při instalaci automaticky připojí program Amos (metodologie SEM – modelování strukturních rovnic).

Větší část této knihy (Část 1, Část 3, Apendixy) je informativní nejen pro uživatele Base, ale i pro uživatele samostatných modulů. Tyto části jsou společné všem modulům. Navíc ovládání procedur v jednotlivých modulech je založeno na stejném principu, a tak postupy statistických procedur popsané v této knize mohou sloužit jako vzory pro většinu procedur všech modulů.

Jádro systému, **IBM Statistics Developer**, je samostatným modulem, obsahujícím všechny obslužné procedury v *Base*. Neobsahuje však statistické procedury, ale jen postupy úprav a manipulace souborů a výstupní editor se všemi jeho funkcemi. Je otevřený k napojení jiných programů, běžně se používá např. jako vhodný základ pro práci s R, neboť jsou tu rychle k dispozici úpravy dat i výstupů, ke kterým lze připojit statistické procedury vytvořené v R. Poskytuje tedy pro vývoj vlastního systému to, co je v běžném programování nejpracnější a trvá nejdelší dobu. Obdobně výhodná spolupráce je k dispozici oblíbeným programovacím jazykem Python.

Editor výstupů, **Smartreader**, je k dispozici bezplatně a může být instalován kdekoliv mimo vlastní systém. Výstupy z programu tak mohou být přenášeny uživatelům výsledků, kteří je mohou nejen číst, ale i editovat v plném rozsahu, aniž by měli nainstalován systém.

Jen několik modulů je funkčních jen v napojení na jiné procedury: **Exact Tests, Bootstrap**, část modulu **Missing Values**.

Program **IBM SPSS Statistics** je ve velké většině případů používán pouze lokálně, všechny výpočty probíhají na počítači, kde je program nainstalován. Při zpracování velkého objemu dat je výhodnější použít architekturu klient-server. V rámci této architektury pak všechny výpočty probíhají na straně serveru. Uživatel se připojuje k serveru přes svoji lokální instalaci programu. Po připojení k **IBM SPSS Statistics Serveru** má uživatel k dispozici moduly podle licence svého lokálního programu a prostředí programu je stejné jako u lokální instalace.

IBM SPSS Statistics Server se instaluje na serverový operační systém a hardware, který má typicky vyšší výpočetní výkon, rychlejší přístup k datům a další vlastnosti zajišťující vyšší bezpečnost dat a důkladnější zálohování.

Používání serveru má hlavně následují výhody:

- vyšší výpočetní kapacita hardwaru a paralelní výpočty serverové verze,
- fyzická blízkost zdrojů dat v databázích a výpočetního jádra, minimalizace provozu sítě,

- algoritmy optimalizované pro načítání dat z databází, částečné zpravování dat přímo v databázi (*pushback*),
- rozšíření algoritmů o *naivní bayesovské klasifikátory* a nástroj výběru vhodných vstupních proměnných do modelů,
- využití zabezpečení serverového operačního systému,
- dávkové zpracování dat v plánovaných úlohách.

Otevřenost systému

Velkou uživatelskou předností systému je jeho **otevřenost**, a to v mnoha směrech:

a) Přímo přebírá soubory nejen svého nativního typu *.sav*, ale i *.xls*, *.xlsx*, *.dbf* a mnoho dalších a také v různých formátech soubory ukládá.
b) Přebírá data ze všech databází, ke kterým je k dispozici napojení ODBC. Velmi důležitou funkcí, otevírající nové zásadní aplikace, je spolupráce s programem *Cognos*.
c) Skripty a makra systému vytvářejí samostatné procedury nebo zpracují výstupní tabulky do uživatelem specifikované formy pomocí jazyku Python.
d) Můžeme k němu napojovat vlastní programy a procedury přímo jako součást systému v jazyku R, Python či jiných programovacích jazycích.
e) Napojuje se přímo i na jiné, speciální samostatné programy, např. na **IBM SPSS Amos**, a to nejen pro souběh či na doplnění probíhajících analýz, ale také jako obslužná funkce datových úprav a přípravy souborů pro aplikace těchto speciálních programů.
f) Ve spojení s *.NET* vytváří uzavřené samostatné aplikace.

Uživatelská příjemnost ('user friendly program')

Uživatelský komfort je velkou předností programu. Projevuje se mnoha aspekty:

- Řízení pomocí menu, nabídkových oken a klávesových zkratek je návodné a přehledné, uživatel je veden nabídkami k volbě zadání. Jde nejen o uživatelské pohodlí, ale i o rychlost, flexibilitu a možnosti rychle opravit chybná či nepřesná zadání.
- Uživatel se může rozhodnout, zda chce pracovat s nabídkovými okny nebo s jednoduchým syntaktickým jazykem, který má mnemotechnickou formu a je snadno zapamatovatelný zapisuje se do samostatného editoru s podrobnou podporou. Připravené instrukce lze uložit, opakovaně použít, snadno měnit a doplňovat jejich parametry a ve Windows automaticky spouštět na aktualizovaných datech. Instrukce syntaktického jazyka lze generovat i z nabídkových oken.
- Jednoduché ovládání a jednoduché a přímé přechody mezi jednotlivými kroky a etapami procesu zpracování.
- Přebírá data z Excelu, .dBase, textových formátů a mnoha jiných formátů; pomocí bezplatně stažitelných ovladačů ODBC také z běžných databází.

- Během statistické analýzy lze otevírat (ze všech dostupných formátů), kopírovat a také jako výsledky procedur programu odvozovat tolik datových souborů, kolik je třeba, a střídavě mezi nimi přecházet, pracovat s nimi, napojovat je a redukovat je podle potřeby.
- Obsahuje techniky organizace dat potřebné k analýze dat a k úpravám datových struktur vhodných pro analýzu – navíc se k těmto úpravám lze vracet kdykoliv v průběhu analýzy.
- Flexibilní práce s pracovními i prezentačními tabulkami a grafy, práce s několika výstupními okny, do nichž lze střídavě ukládat výsledky podle potřeb, a tím je již v průběhu analýzy třídit.
- Dokumentace celého procesu v žurnálu a ve výstupním okně (volitelný přímý záznam v textovém okně a v dokumentačním okně procedury).

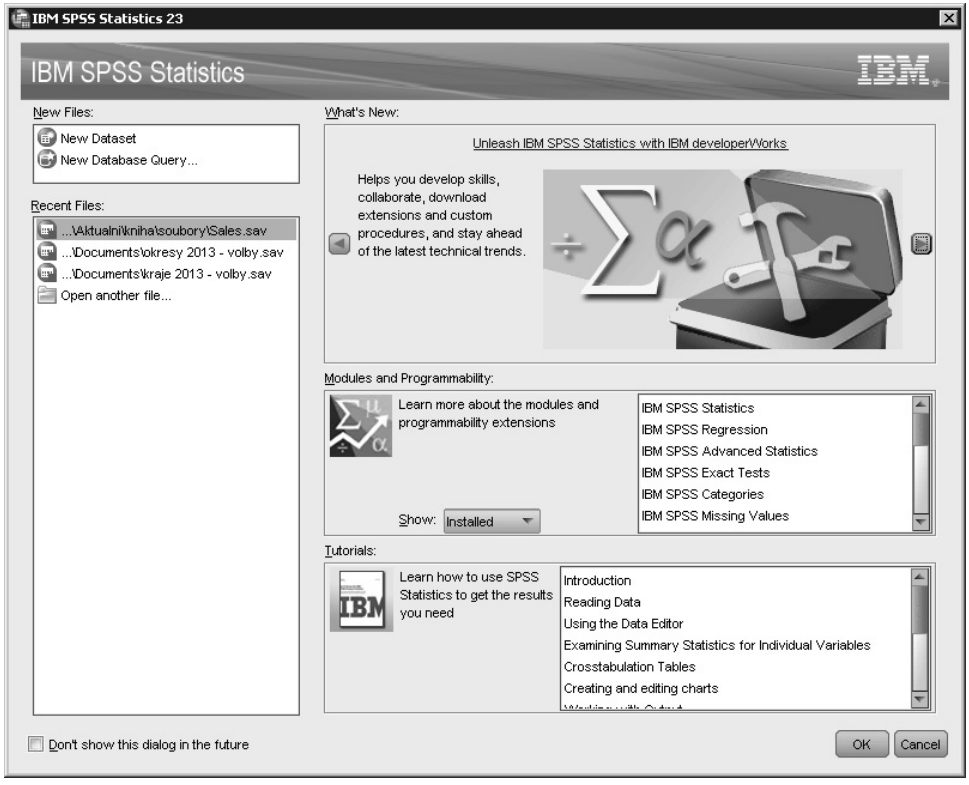
Uživatelská příjemnost má ve svém důsledku velmi podstatný důsledek, protože díky ní uživatel snadno upravuje data, rychle kontroluje průběžné výsledky i ověřuje předpoklady a provádí modifikace a korekce nastavení. Podmiňuje tak rychlou a efektivní cestu k závěrům a šetří čas i zbytečné mezikroky. Nevyžaduje žádné programátorské znalosti ani nutnost pamatovat si formální postupy a přísná pravidla zadávání.

Z uvedených vlastností je také zřejmé, že systém je vhodný pro nejrůznější typy analýz a zpracovatelských procesů. Z obsahu analytických procedur bude také vidět, že s ním může pracovat jak uživatel bez statistických znalostí, který vytváří reporty, tak statisticky poučený analytik, který využívá základní výstupy metod pro datové závěry, i profesionální matematický statistik vyžadující detailní obsluhu a nuance metod, schopný využít jemností modelů pro sofistikované závěry.

Otevřeme program

Po otevření programu (např. kliknutím na ikonu **IBM SPSS Statistics** na ploše počítače nebo na soubor *.sav*) se objeví datová tabulka. Ta je prázdná nebo zaplněná (podle způsobu otevření). V prvním případě se otevře vstupní nabídkové okno. Využijeme jej pro otevření žádaného souboru – buď jednoho z posledně použitých, nebo jej vyhledáme ve složkách počítače (**Open another file**). Vstupní nabídku lze zrušit volbou v levém dolním rohu anebo znovu vyžádat a opět otevřít v menu **`File – Welcome Dialog...`**

Program otevřel dvě okna v záložkách **Data View** a **Variable View**.



Obrázek 1 Vstupní nabídkové okno – poslední pracovní použité uložené soubory, otevření nových datasetů, tutoriály a informace o programu

Data View je tabulka, která je prázdná nebo zobrazuje data aktivního souboru. Zobrazuje data v původních kódech a číslech nebo zobrazí názvy kódů podle určeného předpisu (číselníku). Lze ji editovat podle potřeby či požadavku analytika (viz kapitola 2).

Variable View je tabulka, která určuje vlastnosti proměnných. Tyto vlastnosti lze kdykoliv upravovat či zrušit nebo zavést (viz kapitola 3).

Ovládání programu

Ovládání programu, jak bylo uvedeno výše, je jednoduché, obdobné tomu, čemu jsme zvyklí i z jiných programů každodenní práce. Je řízeno *nabídkovým menu, nabídkovými okny, ikonami,* a *klávesovými zkratkami.* Souběžně s nabídkovým systémem je k dispozici také *mnemotechnický uživatelský zadávací jazyk, syntaxe*. Uživatel se rozhoduje sám, zda bude používat jedno či druhé či oba způsoby v kombinaci.

Nabídkový systém je založen na přehledných *nabídkových záložkách*, které třídí funkce programu dle jejich role a na postupných *zadávacích* nebo *nabídkových oknech*, jejichž struktura odpovídá

danému úkolu, jeho složitosti a jeho parametrům. Práce s nabídkovými okny odpovídá průběžnému rychlému procesu analýzy dat, modifikacím dat podle vývoje úlohy, bezprostředním reakcím na výsledky a opravám nevhodného či chybného zadání. Otevírá také možnost operativních průběžných změn v datovém souboru v procesu analýzy. Vlastní procedury, skripty a připojené programy mohou být reprezentovány ikonami, které si uživatel vytvoří. Kromě standardních tradičních oken jsou v posledních verzích zařazována také speciální okna pro specializované procedury či moduly a pro automatizované postupy.

Syntaxe má výhodu v přípravě dávkového výpočtu, možnosti uložit zadání a snadno měnit jeho parametry, zkrácení postupu při zadávání opakovaných úkolů, a vytvoření podkladu pro automatické jednorázové či opakované spouštění programu ve Windows. Syntaxe obsahuje širší možnosti než okna, neboť mnoho analytických a manipulačních kroků a voleb používáme zřídka a jejich zařazení do oken by komplikovalo přehlednost oken, a tím běžnou standardní práci.

Příkazy syntaxe zapisujeme do zvláštního okna, které proces ulehčuje řadou podpůrných funkcí. Uložený syntaktický proud příkazů používá označení s koncovkou *.sps*. Příkazy, které jsou ekvivalentní konkrétní volbě v nabídkových oknech, lze automaticky generovat tlačítkem **Paste** (a poté případně uložit nebo modifikovat). Syntaktický uživatelský jazyk *de facto* do praxe ovládání analytických programů zavedli jako první autoři SPSS už v šedesátých letech minulého století. V té době, kdy neexistovaly možnosti dialogového zadávání, tato inovace znamenala průlom do použití statistiky, protože uživatelé přestali být závislí na složitém zadávacím postupu jednotlivých programovacích jazyků a mohli si své výpočetní běhy připravovat sami.

Jednoduchá a výstižná mnemotechnika a struktura příkazů byla důvodem velké a rychlé popularity systému SPSS mezi uživateli, vytvořila základ pojmu „uživatelská příjemnost" a otevřela přímou cestu ke statistice pro vědce, výzkumníky, manažery, a to i s naprosto zásadním významem pro výuku, studenty i učitele. Princip syntaktického jazyka se nemění po celou dobu vývoje systému SPSS, jazyk je pouze doplňován pro nové procedury.

Pomocí syntaxe lze zadat řadu aktivit, které by pro své nefrekventované používání nebo pro složitost zadání komplikovaly jednoduché postupy oken. V této knize se soustřeďujeme na práci se zadávacími okny nabídky. Omezení místa a objemnost látky nedovoluje zabývat se podrobněji syntaktickým jazykem SPSS, jehož základnímu popisu věnujeme Apendix A. Podrobný popis jednotlivých příkazů se otevře v záložce základních oken systému **`Help – Command Syntax Reference`**.

Kroky v postupu práce: data, analýza, výstupy

Každý modul se skládá z procedur poskytujících určité specifické aktivity. Role jednotlivých modulů i jejich procedur v zapojení do procesu datového zpracování se od sebe liší. Tyto role se podřizují třem obecným funkcím programu:

- přípravě dat na analýzu (viz Část 1)
- analytickému zpracování dat (viz Část 2)
- práci s výstupními tabulkami a grafy (viz Část 3)

Kromě toho máme v programu k dispozici řadu funkcí, které usnadňují postup a urychlují průběžnou práci.

Příprava dat a operace s nimi před analýzou a při ní se týká souboru jako celku, případů (řádků datové matice) a proměnných (sloupců datové matice). **IBM SPSS Statistics** poskytuje velmi bohaté portfolio technik pro tuto etapu. Většina z nich je zahrnuta v modulu *Base*, specifické postupy jsou ale uloženy v modulech *Data Preparation* a *Missing Values.* Také modul *Complex Samples* má částečně přípravný charakter.

Primárním cílem systému je ovšem poskytnout statistickou podporu zpracování informací a získání výsledků pro následné využití v praxi. Proto zde nalezneme všechny běžně používané statistické metody pro analýzu dat a její závěry, a to jak na základní, tak i na pokročilé úrovni. Vzhledem k otevřenosti systému (výhodné využití jazyka R, možnost napojení vnějších nezávislých programů, práce s Pythonem a *.NET*) tak může být použit pro rutinní praxi i pro velmi speciální a sofistikované analýzy za použití metod, které v systému přímo zahrnuty nejsou, ale návazně vystupují v procesu. Typickým případem je modelování kauzálních vztahů přechodem do programu **IBM SPSS Amos**.

Vizualizace výsledků a tabulkové výstupy jak pro pracovní průběžné cíle, tak pro prezentaci výsledků jsou flexibilní a využívají předvolené šablony nebo vlastní vytvořené šablony.

Menu nabídkové lišty

Menu nabídkové lišty a ikony se liší podle typu souboru *sav* (data, výstupy, syntaxe). Záložky třídí procedury podle typu funkcionality v pracovním procesu.

V datovém editoru má hlavní lišta záložky pro všechny etapy práce:

Tabulka 2 Záložky programu v oknech Data View a Variable View

Název záložky	Data View
File	převzetí a ukládání souborů, tisk
Edit	editace oken
View	úpravy okna
Data	úpravy dat, kontrola kvality
Transform	konstrukce nových a úpravy původních proměnných
Analyze	procedury zpracování dat
Direct Marketing	procedury aplikačního modulu
Graphs	grafické prostředky systému
Utilities	zavádění maker, procedur a skriptů, podpůrné funkce
Add-ons	informace o modulech a dalších programech rodiny IBM SPSS
Window	použití oken
Help	popisy funkcí, tutoriál, algoritmy, syntaxe, případové studie, práce s R a Pythonem

Jednotlivé záložky, především **Analyze**, jsou naplněny podle rozsahu instalace modulů. Záložka **Direct Marketing** odpovídá celá jednomu modulu. Vytváří-li uživatel své vlastní procedury či makra, mohou jím být zavedeny další specifické záložky. Procedury jednotlivých záložek jsou

vypsány v Apendixech D (nativní procedury systému), E (procedury založené na jazyce Python) a F (procedury v jazyce R)

Ikony

Sada ikon se v obou vstupních oknech, ve výstupním okně a syntaktickém editoru liší. Průnikem jsou běžné akce týkající se univerzálních kroků v procesu, jako jsou: ukládání, tisk, otevření souboru, rušení akce a návrat ke zrušenému, vyhledávání, přechody v rámci souboru, vkládání případů a proměnných, pouštění skriptů. V jednotlivých oknech pak jsou přidány ikony akcí specifických pro toto okno. Název ikony (její funkce) se objeví, najedeme-li na ni myší. Jednotlivé ikony jsou aktivované jen tehdy, mají-li smysl.

V **Data View** a ve **Variable View** je to navíc například vážení, rozdělení souboru a výběry podsouborů. Pro označenou proměnnou (v každém z obou oken) ikona **Run descriptive statistics** spočte základní míry. V **Data View** je navíc důležitá provozní ikona **Value Labels**, která v datové matici přepíná kódy na názvy a naopak (funkce toggle), takže pomáhá k rychlé orientaci v řádku či sloupci.

Ve výstupním okně (**Viewer**) jsou záložky stejné, ikony se váží k editaci výstupu, resp. k analýze výstupních dat pomocí skriptů. Jsou to akce otevírání objektů, skrývání a znovuotevření objektů, funkce zavádění autoskriptů. V tomto okně ale můžeme mít zavedeny ikony pro vyvolání skriptů, máme-li takové připraveny. Vlastní ikony mají editační okna grafů a okna pivotních tabulek.

V editoru syntaxe jsou umístěny ikony pro editaci příkazů a přímé vyvolání pomoci pro označený příkaz.

Velmi užitečnou interakční ikonou ve všech oknech je **Dialog Recall** (**Recall recently used dialogs**), ve které je seznam posledních použitých procedur a po jejímž potvrzení se potvrzením vybrané procedury dostaneme přímo k poslednímu zadání pro daný dataset. Tato ikona velmi zrychluje analýzu a podporuje „rozhovor" analytika s daty jednak v procesu upřesňování úlohy, jednak při chybných zadáních.

Skripty, makra, procedury uživatelů

Standardní výstupy z jednotlivých analýz mohou být automaticky nebo volitelně obměněny pomocí skriptů – (mini)programů v jazyce Python, které buď výstupní tabulky modifikují, editují a přeorganizují, nebo na základě získaných výsledků dopočítají nové statistiky, aplikují na nich další metody, které ve standardním výstupu nejsou, a vytvářejí nové, odvozené tabulky. Tyto skripty připravuje nebo přebírá uživatel.

Skripty jsou velmi užitečné doplňky základních výstupů. Doplňují analýzu, zpřehledňují výstupy podle vkusu uživatele, a to buď:

- na manuální vyžádání vyhledáním ve složce **`Utilities – Run Script...`** , nebo
- automaticky při výstupu – *autoscript*.

Tyto programy lze vybavit nabídkovými okny podle přání a variant zpracování. Na lištu výstupového okna **Viewer** můžeme umístit vlastní připravenou ikonu pro přímé vyvolání skriptu na označený výstup.

Skripty se typicky vytvářejí na podbarvení tabulky nebo zvýraznění hodnot, na zjednodušení tabulky, dopočítání testů významnosti, které nejsou zahrnuty v proceduře, sumarizace výsledků z několika tabulek. Skripty si vytvářejí uživatelé sami, některé skripty přicházejí se systémem a existuje mnoho veřejně dostupných zdrojů s možností stáhnout si je a používat (jedním z volných zdrojů jsou webové stránky autorů, *www.acrea.cz*, kde lze nalézt řadu praktických skriptů pro analytickou práci uživatelů). Autoskripty zavádíme pro jednotlivé procedury a typy výstupů proto, abychom dostali přímo takový tvar výstupů, jaký nám vyhovuje lépe, než jak jej předvolili autoři systému. Úpravu pak nemusíme provádět jednotlivě.

Systém **IBM SPSS Statistics** má také svůj vlastní maticový jazyk, ve kterém můžeme zadávat různé algoritmy a vytvářet tak speciální procedury pro analýzu dat bez použití vnějších programovacích prostředků.

Procedury vnějšího původu (programované v R, v Pythonu nebo uzavřené programy) můžeme připojit do menu a pracovat s nimi stejně jako s nativními procedurami.

Vývoj systému

Systém přichází každý rok s novou rozšířenou verzí, jsou připojovány nové procedury, někdy celý nový modul, rozšiřují se jak postupy analytické, tak postupy úpravy dat i editace. Ve verzi 23 sytému byla například do modulu **Base** připojena zásadní novinka – procedura časově-prostorových analýz a predikcí (z důvodů místa není v této knize popisována). Kromě těchto viditelných aspektů jsou to ale i ty, které zvnějšku nevidíme, pocítíme je až při analytické práci samotné – zvyšování rychlosti, přesnosti a spolehlivosti zaváděním nových algoritmů a či přizpůsobení se k vývoji operačních systémů a reakce na prudce se zvyšující objemy datových zdrojů.

Systém reaguje na vývoj hardwarových i softwarových technologií, na rozmanitost i rozsahy informačních kontextů a na nutnost získávat precizní podklady rychle a komplexně. Je flexibilní k požadavkům analytiků a otevírá se stále více zapojování vnějších programových prostředků. Schopností vstřebávat snadno vnější příspěvky (R, Python) ovšem podstatně zrychluje i rozšiřování portfolia své statistické nabídky a také zvyšuje potenci participace uživatelů v procesu vývoje.

ČÁST

PŘÍPRAVA DAT

V této části:

- **KAPITOLA 1** – Soubory
- **KAPITOLA 2** – Případy
- **KAPITOLA 3** – Proměnné

Před analýzou dat

Příprava datového souboru je nejpracnější etapou analytické práce. Data zapisujeme nebo přebíráme, čistíme, prověřujeme jejich kvalitu, upravujeme pro analýzu, vytváříme nové proměnné a podnikáme kroky zajišťující jednoduchou, rychlou a efektivní práci v dalších etapách procesu. Funkce, které program poskytuje, zjednodušují nejen přípravné práce, ale také jakékoliv nutné či vhodné změny v průběhu analýzy.

Datové zdroje předpokládají přípravné, modifikační a kontrolní činnosti, které se dělí na tři skupiny – každou z nich popisuje jedna kapitola:

- *Kap. 1 Soubory* – úprava souboru jako celku, vlastnosti celé datové matice
- *Kap. 2 Případy* – jednotlivé případy – práce s případy, řádky datové matice
- *Kap. 3 Proměnné* – příprava proměnných, sloupců datové matice

Výsledky těchto změn platí tak dlouho, dokud nejsou zrušeny či přeměněny jinými změnami. Lze je samozřejmě i uložit do používaného souboru nebo do souboru nového.

Modul **IBM SPSS Statistics Base** podporuje přípravné fáze velkým počtem procedur (další speciální procedury pro tuto etapu jsou obsahem modulu **IBM SPSS Statistics Data Preparation**).

Základní úkoly přípravných i průběžných zásahů do datového souboru jsou:

a) vybavit soubor stálou informací pro snadnou aplikaci, orientaci a korektní používání proměnných;
b) identifikovat případy nebo skupiny případů, které do souboru pro daný účel nepatří (chyby při záznamu, nesourodé případy, duplikáty), a opravit je nebo vyloučit;
c) zbavit soubor chyb a identifikovat vynechávané hodnoty;
d) změnit původní a/nebo vytvořit nové proměnné transformací;
e) vytvářet účelové podsoubory;
f) spojovat a agregovat soubory.

V této části uvádíme speciální procedury pro tento účel, které jsou obsahem modulu **Base**. S daty, s jejich úpravami a doplňováním pracujeme v průběhu celého analytického procesu. Vybavení souboru můžeme kdykoliv změnit. Kvalitu dat ověřujeme nejen procedurami této části, ale také ve statistických procedurách (Část 2) i pomocí pracovních grafů (Část 3). Procedury Části 2 jsou součástí každého modulu a dají se v jeho rámci využívat i bez přítomnosti modulu **Base**.

KAPITOLA 1

Soubory

V této kapitole:

- Manuální zápis dat do souboru
- Převzetí datového souboru do programu
- Vybavení souboru – Variable View
- Datasety
- Transpozice
- Restrukturace
- Spojování souborů
- Agregace případů

Soubory pro statistickou práci jsou vždy připraveny ve tvaru datové matice – obdélníkové tabulky, jejíž řádky zpravidla odpovídají případům a sloupce proměnným. Datovou matici tvoříme či přebíráme buď přímo z programu **IBM SPSS Statistics**, nebo z jiných forem zápisu, jako jsou relační databáze, textové soubory či tabulkové procesory. Při analýze se předpokládá, že pracovní soubory jsou již připravené ve tvaru datové matice.

Práce se soubory zahrnuje:

a) vytvoření nebo převzetí pracovních souborů/datasetů
b) vybavení souborů pro analýzu i pro vhodné výstupy
c) transpozice souborů, tj. záměna řádků a sloupců v jejich analytické roli
d) restrukturace souborů na vhodný analytický tvar (částečná transpozice)
e) spojování souborů
f) agregování souborů
g) rozdělení souboru na části pro paralelní výpočty

Operace se soubory jsou podstatnou částí analytické práce. Zpracování dat je podstatně ulehčeno dobrým vybavením souboru. Některé úlohy předpokládají pro ně nutný či vhodný tvar souboru.

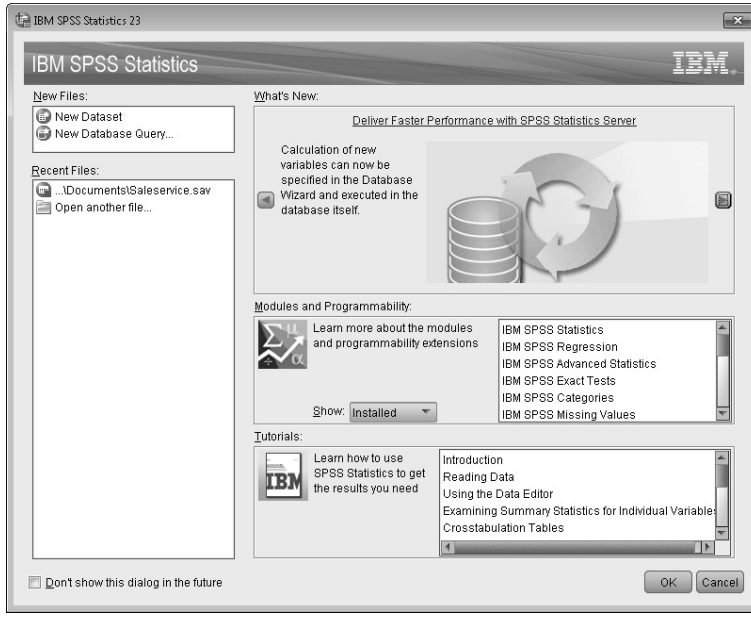
Manuální zápis dat do souboru

Malé soubory dat můžeme zapsat manuálně přímo jako pracovní soubor do nového prázdného datového okna, tj. do nového tzv. *datasetu*.

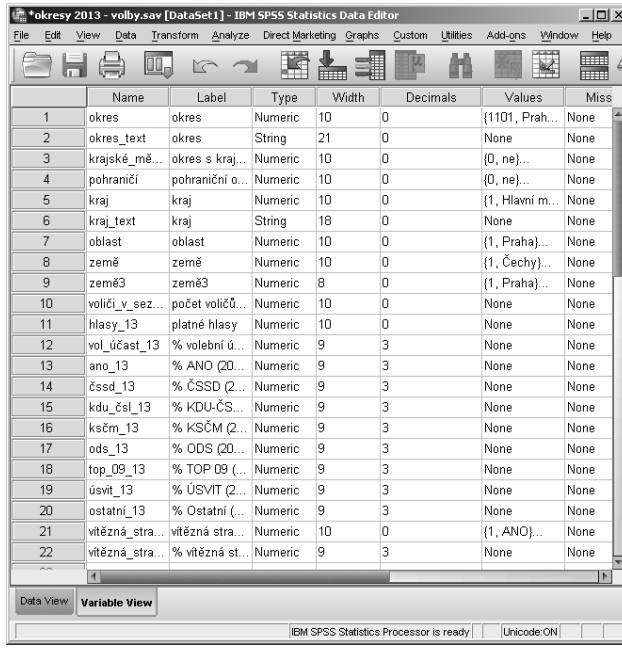
Postup A – při vyvolání programu se otevře vstupní nabídka:

1. otevřeme program,
2. ve vstupní nabídce zvolíme v levém horním okně **New Files** řádek **New Dataset**,
3. záložka **Variable View** otevře okno proměnných, v něm pojmenujeme proměnné (sloupce), určíme jejich vlastnosti,
4. v otevřeném prázdném datovém oknu (**Data View**) se data pro jednotlivé případy (řádky) zapisují do příslušných sloupců, které jsou již pojmenovány,

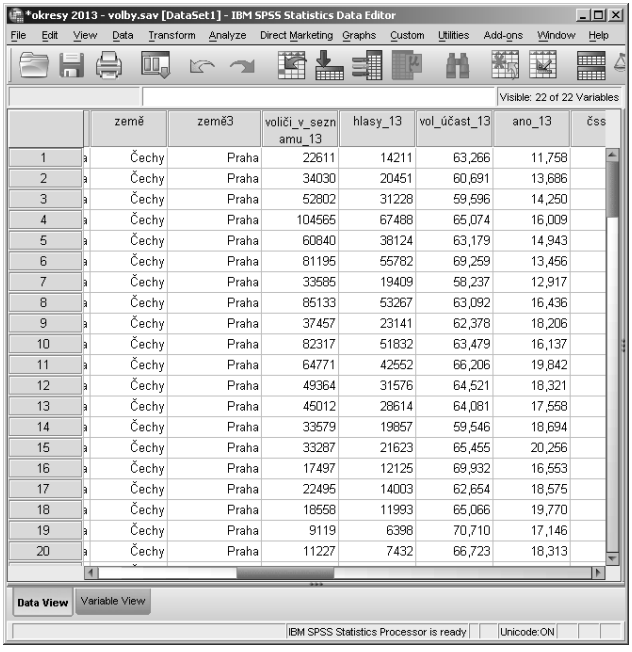
5. nový řádek se otevře při zápisu první hodnoty.



Obrázek 1.1 Okno vstupní nabídky při otevření programu



Obrázek 1.2 Okno záložky ***Variable View*** – vybavení proměnných



Obrázek 1.3 Datové okno s pořízenými hodnotami

Postup B – z hlavního menu kdykoliv v průběhu práce:

1. otevřeme program
2. zvolíme nabídku **`File – New – Data`**
3. ve **Variable View** pojmenujeme proměnné (sloupce), určíme jejich vlastnosti
4. v otevřeném prázdném datovém oknu (**Data View**) se data pro jednotlivé případy (řádky) zapisují do příslušných sloupců, které jsou již pojmenovány
5. nový řádek se otevře při zápisu první hodnoty.

Kroky 4 a 5 mohou být nahrazeny kopírováním dat např. z Excelu.

V obou případech se nový soubor nazve automaticky *Dataset* s pořadovým číslem. Přejmenujeme jej ve **`File – Rename Dataset`**. Zde se otevře okénko, v němž se žádané jméno zapíše.