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

ČÁST III

VÝSTUPY A JEJICH ÚPRAVY

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).

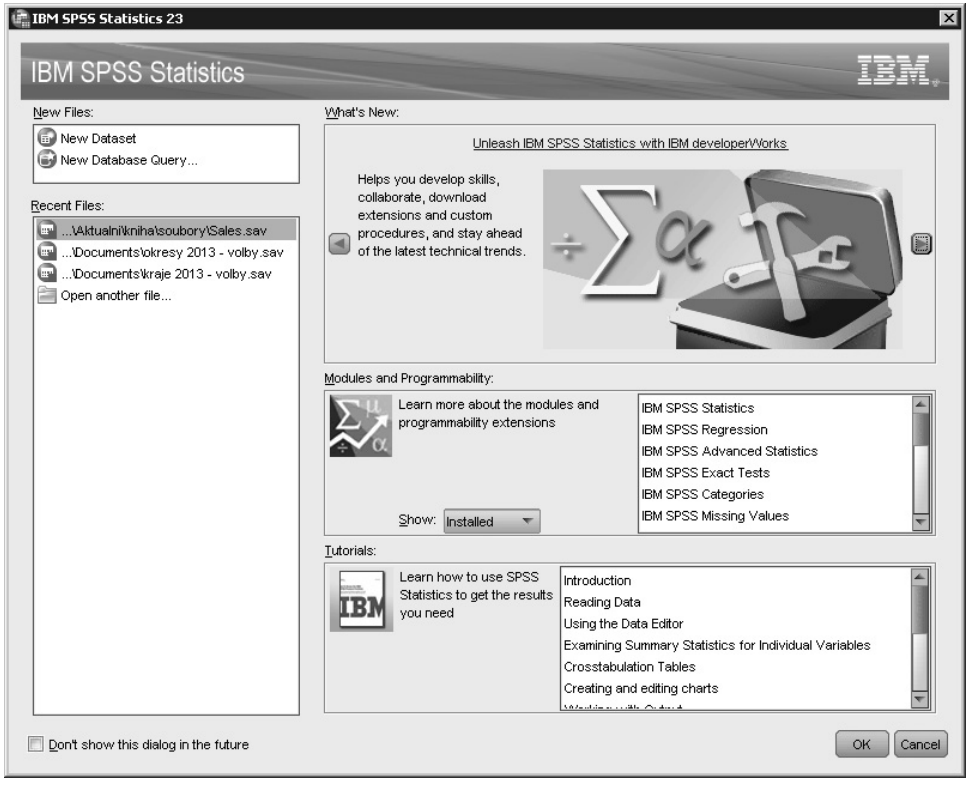
Uživatelská příjemnost má ve svém důsledku velmi podstatný důsledek, protože díky ní uživatel snadno upravuje data, rychle kontroluje průběžné výsledky i ověřuje předpoklady a provádí modifikace a korekce nastavení. Podmiňuje tak rychlou a efektivní cestu k závěrům a šetří čas i zbytečné mezikroky. Nevyžaduje 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.

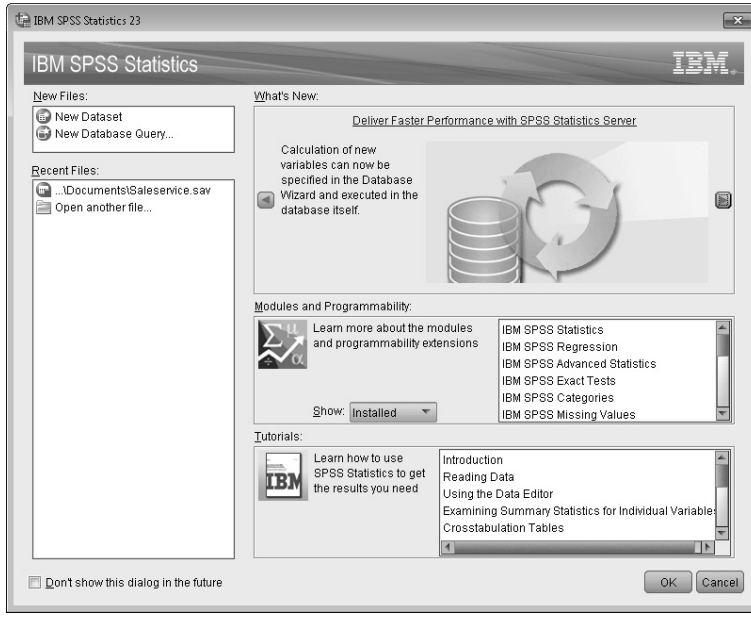
Manuální zápis dat do souboru

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

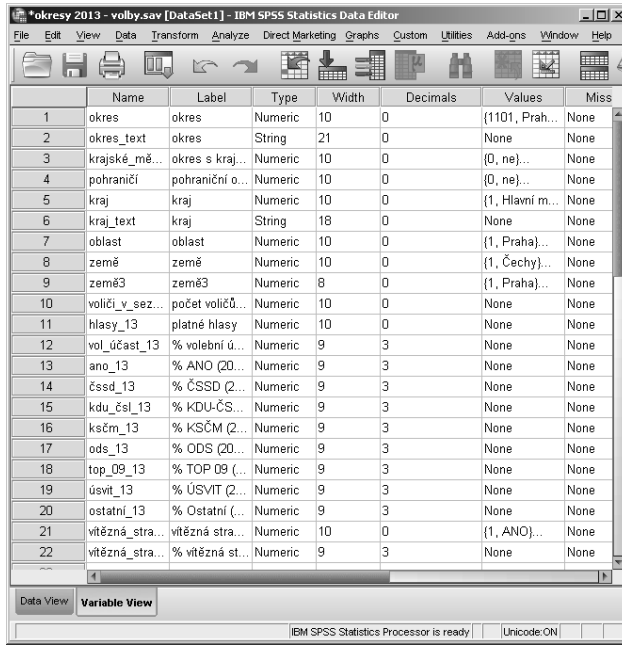
Postup A – při vyvolání programu se otevře vstupní nabídka:

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,

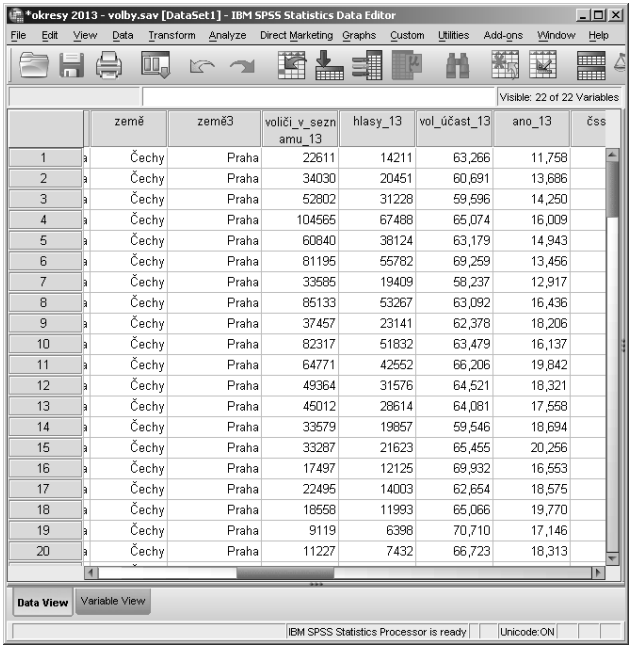
5. nový řádek se otevře při zápisu první hodnoty.



Obrázek 1.1 Okno vstupní nabídky při otevření programu



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



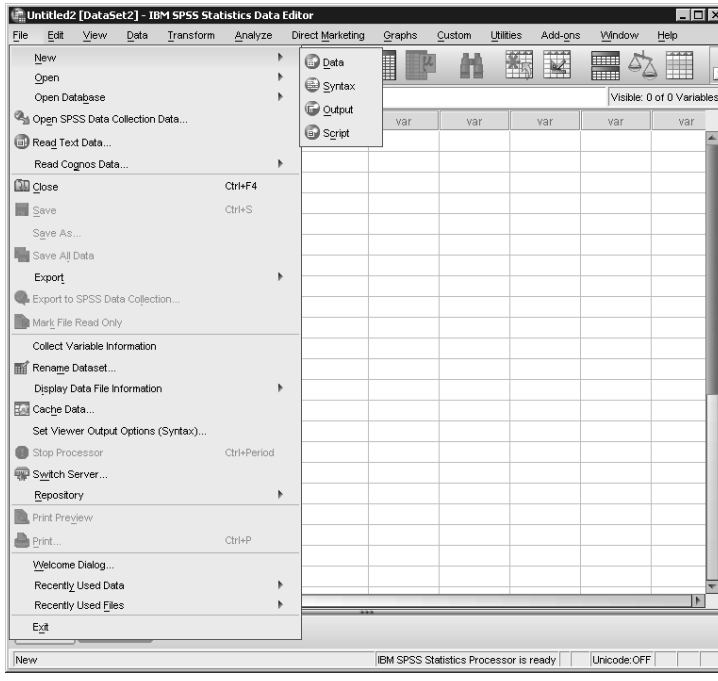
Obrázek 1.3 Datové okno s pořízenými hodnotami

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.



Obrázek 1.4 Zavedení nového souboru *File – New – Data*

Při pojmenovávání proměnných se automaticky zavede číselný formát F8.2 pro datovou matici (8 značí šířku čísla a 2 je počet zobrazovaných desetinných míst) – počet v souboru zapsaných a používaných desetinných míst může být jiný (!). Jde-li o textovou proměnnou, předvolená délka textu je 8. Předvolené parametry můžeme změnit podle potřeby.

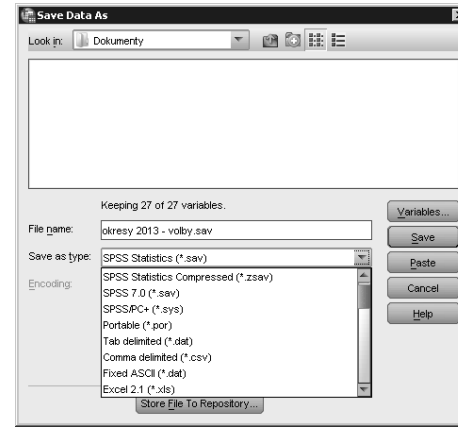
Z hlediska metodiky statistické práce zdůrazňujeme, že všechny nově pořizované soubory musí být – pro zajištění kvality dat i výsledků – nutně zkontrolovány v plném rozsahu všech případů a proměnných.

Soubor se stane aktivním již v průběhu zapisování, lze jej zpracovávat a uložit.

Nový soubor ukládáme tak, že:

ve volbě **`File – Save as...`** nalezneme příslušnou složku, zapíšeme název do řádku ***File name*** a určíme typ v řádku ***Save as file***. Předvolbou je nativní typ *.sav*, lze jej však změnit podle nabídky.

Možností tu je také přiřadit heslo k otevírání souboru zatržením volby ***Encrypt file with password***.



Obrázek 1.5 Ukládání souboru: *File – Save as*

Převzetí datového souboru do programu

V běžné praxi jsou soubory již pořízené a uložené buď ve formátu *.sav*, nebo v jiných běžných formátech.

Převzetí souborů je vedeno několika způsoby:

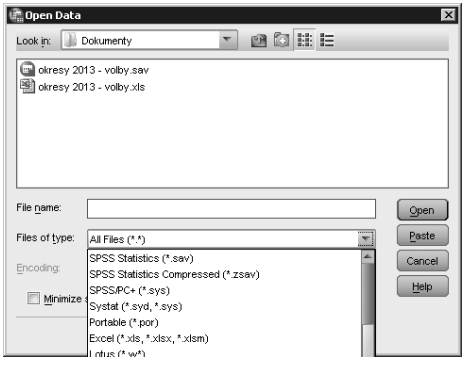
a) přímé převzetí datové matice z některého formátu *.sav*:

b) základní formát *.sav*, komprimovaný formát *.zsav*, též formáty z období DOS *.sys* (formát dosovského souboru) a *.por* (přenosový formát);
   přímé převzetí datové matice z jiných vybraných formátů:
   - soubory Excelu (*.xls*, *.xlsx*, *.xlsm*)
   - textové soubory *(.txt*, *.dat*, *.csv*, *.tab*)
   - soubory dBase (*.dbf*)
   - soubory jiných statistických programů – Stata (*.dta*), SAS (*.sas7bdat*, *.sd7*, *.sd2*, *.ssd01*, *.ssd04*, *.xpt*), Systat (*.sys*, *.syd*),
   - Sylk (*.slk*)
   - Lotus (.*w*)

c) převzetí dat z různých relačních databází pomocí ODBC;

d) EXCEL a ACCESS jsou předvoleny, při dodávce programu jsou k dispozici další ODBC; postup kopíruje posloupnost nabídek;

e) soubory programu Cognos.
   Po otevření souboru v pracovním režimu používáme datový formát *sav*.



Obrázek 1.6 Převzetí souboru – specifikace formátu

Program může mít současně otevřených několik pracovních souborů, ať už jsou převzaty jako datová matice, vytvořeny v průběhu práce, či vytvořeny manuální volbou. Ty jsou nazývány *datasety*, dostávají své jméno a mohou být uloženy jako *.sav* nebo jiný typ výstupového formátu, který je k dispozici v nabídce **`File – Save as...`**

Samotné přímé převzetí souborů *.sav* je možné několika způsoby:

a) Otevřeme program a ve vstupní nabídce volíme v okně ***Recent Files*** ze seznamu předchozích použitých souborů nebo vyhledáme soubor v **Open another file …**

b) Na začátku – i kdykoliv během práce – můžeme otevřít soubor cestou ***`File – Open – Data ...`*** **– *vyhledat soubor …***

c) Předchozí soubory jsou uvedeny v menu **`File – Recently Used Data ...`** (jejich počet v rozmezí nula až deset je volitelný v **`Edit – Options – File Locations`** – v okně ***Number of Recently Used Files to List***)

d) Dvojí poklepání na soubory s nativní koncovkou *.sav*

e) Přenesením, levou myší, ikony souboru *.sav* nebo souboru, který **IBM SPSS Statistics** čte přímo na ikonu jeho zástupce

f) Přenesením, levou myší, ikony souboru *.sav* nebo souboru, který **IBM SPSS Statistics** čte přímo, kamkoliv do pole otevřeného programu

Postupy e) a f) lze aplikovat nejen na soubory *.sav*, ale např. i na soubory MS Excel.

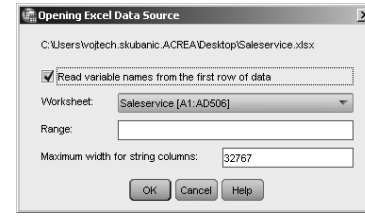
Program s prázdným datovým oknem otevřeme ikonou nebo také potvrzením ze seznamů v obslužných programech Windows či přímo vyvoláním *stats.exe* ze složky *IBM/SPSS/Statistics/23* (resp. číslo instalované verze).

Zcela obdobně se otevřou soubory syntaxe (*.sps*) a výstupů (*.spv*).

Soubory *.sav* se otevřou s celou uloženou výbavou v **Data View**.

Jako příklad uvedeme časté přebírání souborů z jedné tabulky Excelu postupem ad b). Postup je obdobný jako při otevření *.sav*:

Po volbě **`File – Open – Data`** přepneme v nabídkovém řádku ***Files of type*** na volbu ***Excel (*.xls, *.xlsx *.xlsm),*** nalezneme žádaný soubor a potvrdíme. Otevře se specifikační okno ***Opening Excel Data Source,*** které vyžaduje určení listu v Excelu (***Worksheet***). Pokud nejsou data umístěna v levém horním rohu, je nutno určit umístění datového obdélníku (***Range***). Datový obdélník může či nemusí obsahovat v prvním řádku názvy sloupců. Tento fakt musíme určit zaškrtnutím v poli ***Read variable names from the first row of data***. Mají-li sloupce v prvním řádku jména, jsou převzaty jako názvy proměnných v pracovním souboru. Nejsou-li jména určena, proměnné v souboru *.sav* jsou nazvány *V1*, *V2* … Typ proměnné je odvozen z prvního řádku dat.



Obrázek 1.7 Specifikace pro převzetí souboru MS Excelu

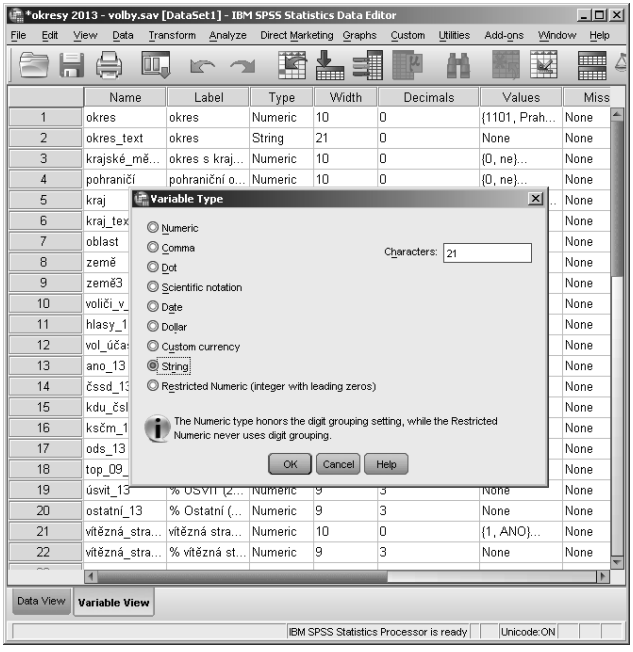
Ze souboru MS Excel tedy přenášíme jen název proměnné a typ proměnné. Musíme dát ale pozor na správné určení prvního řádku – neurčíme-li jej jako řádek s názvy a on přitom názvy obsahuje, program převezme řádek jako datový a určí všechny proměnné jako textové (***String***). Pracovní soubor bude mít počítačem určené jméno *Dataset*. To změníme následným uložením souboru jako *.sav* (**`File – Save As – ...`**), pojmenováním datasetu (**`File – Rename Dataset`** – ***zápis jména***) nebo obojím.

Vybavení souboru – Variable View

Vybavenost souboru stálými parametry jednotlivých proměnných zajišťuje uživatelský komfort jak při analýze, tak při finální editaci výsledných tabulek a grafů. Proto vybavení souboru věnujeme vysokou pozornost již při převzetí dat. Můžeme je ale měnit kdykoliv během práce.

Soubor typu *.sav* obsahuje dvě části: datovou matici (**Data View**) a tabulku vlastností proměnných (**Variable View**), které se přepínají na základní liště.

Vybavení datové matice v okně **Variable View** podrobnou informací o proměnných (sloupcích souboru) je předností systému **IBM SPSS Statistics**. Každý datový sloupec je charakterizován jednak *popisnou* a jednak *pracovní* informací.



Obrázek 1.8 Okno záložky *Variable View*

Parametry popisu proměnných datové matice určíme a měníme kliknutím na příslušné políčko ve **Variable View**:

- **Name** – jméno proměnné
  - určujeme přímým zápisem
  - je určující pro použití sloupce/proměnné v jakékoliv akci systému
  - je v souboru jen jednou (dvě jména proměnných v jednom souboru systém nepřijme)
  - musí začínat písmenem (nebo speciálním znakem pro speciální roli)
  - jména mohou obsahovat českou diakritiku
  - proměnné začínající znaky $, # a @ mají speciální roli v systému (např. *$Casenum* znamená automatickou proměnnou aktuálního pořadí řádku v souboru, další se týkají data a času, systémově vynechaných hodnot), # jsou pomocné v systému)
  - může mít až 64 libovolných znaků, ale nesmí obsahovat mezery a interpunkční znaménka s výjimkou podtržítka a tečky uvnitř jména
  - jsou vyloučena slova ALL, AND, BY, EQ, GE, LE, LT, GT, NE, NOT, OR, TO a WITH; to jsou klíčová slova v syntaxi a v řízení programu, která mají ve spojení s proměnnými specifický význam (viz Apendix A)

Tip: Proměnné je vhodné pojmenovat číslem záznamu v původním zdroji (např. v dotazníku nebo ve formuláři) nebo mnemotechnicky zkratkou významu proměnné – např. Ot.1. Ot.2 … nebo datnar, titul, vzdělání …

- **Type** – typ informace
  - volíme a specifikujeme v nabídce pole, vybereme typ a jeho formát
  - vyjadřuje pokyn pro počítač, že informace je určitého typu; hlavní typy:
    - ***Numeric*** – číslo, číselně zpracovatelná informace
    - ***String*** – text, textový záznam (do 32 767 bytů)
    - různé tvary numerického záznamu (znak $ na začátku čísla, záznamy s oddělovacími znaky, vědecká notace, celá čísla s předsazenými nulami)
    - ***Custom currency*** – uživatelem volitelný prefix a/nebo sufix; volbu až pěti různých takových formátů provedeme předem v `Edit – Options – Currency`
    - ***Date*** – formáty data a času
  - může být měněn v přípravě či v průběhu analýzy; je ale třeba být opatrný na možnou ztrátu některých hodnot, např. při převodu textově přijatých číselných hodnot s desetinnou tečkou na číselný formát používající desetinnou čárku či naopak, při chybném převodu na formát času apod.

Tip: Při standardním nastavení českých Windows se v IBM SPSS Statistics zobrazuje desetinná tečka jako čárka. Situaci vyřešíme formátem *COMMA*, který naopak pracuje s desetinnou tečkou.

- **Width** – šířka hodnoty proměnné v zobrazení dat (počet cifer, počet písmen) se určí přepisem nebo nabídkou
- **Decimals** – počet zobrazovaných desetinných míst se určí přepisem nebo nabídkou
- **Label** – název proměnné se zapíše přímo
  - popisný text, který reprezentuje proměnnou v tabulkách a grafech
  - text obsahuje libovolné znaky
  - všechny procedury tisknou 40 znaků nebo více (až do 255 znaků)
  - pro anglické názvy je k dispozici kontrola (*spelling*) pro všechny názvy ve sloupci **Label**
- **Values** – názvy kódů (kódový klíč)
  - v nabídkovém okně pole se napíše hodnota kódu (***Value***) a název (***Label***) a tlačítkem ***Add*** se připojí do seznamu (číselníku)
  - při potvrzení řádku kódového předpisu je nabídnuto odstranění (***Remove***), přepíšeme-li název, nabídne se změna (***Change***), přepíšeme-li hodnotu, je nabídnuto přidání nebo změna
  - volitelné názvy jednotlivých hodnot proměnné mají délku až 120 znaků
  - mohou být přiřazeny kterékoliv definované proměnné, jakéhokoliv typu
  - nejsou povinné, a ani nemusí být určeny pro všechny hodnoty
  - mohou být určeny i pro hodnoty, které se v souborech nevyskytnou
  - mohou být jednotlivě přidávány, měněny a odstraňovány kdykoliv během analýzy pomocí nabídky, kterou dostaneme po kliknutí na dané políčko
  - tytéž názvy mohou být přiřazeny více proměnným, lze je kopírovat Ctrl-C/V
  - názvy mohou být stejné pro různé kódy
  - pro anglické názvy je k dispozici kontrola (*spelling*) pro všechny názvy ve sloupci i jednotlivě pro proměnné

Tip: Názvům proměnných věnujte velkou pozornost – jde o podstatnou informací pro čtenáře výstupů a určuje jejich interpretaci. Proto musí být název informačně úplný a přesný, ale zároveň i přehledný, aby tabulku či graf nezahltil.

- **Missing** – hodnoty proměnné, které jsou z analýzy vynechávány; jsou to obvykle chybějící hodnoty, hodnoty, které jsou netypické a pravděpodobně chybné, nebo hodnoty, které chceme dočasně z analýzy vyloučit. Systém pracuje se dvěma druhy vynechávané informace: *systémově vynechávané (system-missing)* a *uživatelem vynechávané (user-missing)*
  - *systémově vynechávaná informace*, v datech označovaná tečkou, vzniká:
    - není-li u definované proměnné určená (zapsána, přejata) hodnota
    - není-li možné provést určenou transformaci proměnné, např. dělení nulou, odmocnina nebo logaritmus záporného čísla
    - prázdná pole při konverzi textové proměnné na numerickou
    - nelze-li provést konverzi v poli
    - převedeme-li hodnoty na systémově vynechaná data příkazem v **`Transform – Recode into Same Variable`** nebo v **`Transform – Recode into Different Variable`** (pozor: po této transformaci se původní údaj ztratí)
  - *uživatelem vynechávané informace:* kromě systémově vynechávaných pozic, značených v datové matici tečkou, můžeme určit i vybrané kódy, které chceme z analýzy vyloučit
    - určíme je v nabídkovém okně po kliknutí na dané políčko
    - k dispozici jsou tři volby:
      - žádné vynechávané hodnoty (předvolba)
      - tři různé hodnoty pro číselné i textové proměnné
      - pro číselné hodnoty – jedna hodnota a uzavřený interval, jehož hranice jsou určeny v polích ***Low*** a ***High***; není-li interval omezen zezdola, zapíšeme *lowest* nebo krátce *lo*, není-li interval omezen seshora, zapíšeme *highest* nebo *hi*
    - volbu pro jednu proměnnou lze kopírovat ***Copy/Paste*** nebo *Ctrl-C/Ctrl-V* pro další proměnné
    - volbu *user-missing* hodnot můžeme měnit kdykoliv během analýzy dat
  - existující systémově vynechaná data můžeme kdykoliv v průběhu analýzy překódovat do zvolených konkrétních čísel nebo textových hodnot a nakládat s nimi běžným způsobem
- **Columns** – šířka sloupce v datové matici
- **Align** – zarovnání hodnot ve sloupci datové matice
- **Measure** – typ proměnné z hlediska jejích vlastností v analýze
  - na rozdíl od **Type**, který se váže na technické zpracování, **Measure** je určeno podle statistických vlastností a aplikací, které je možné u nich provést
  - u číselných proměnných to jsou
    - ***Nominal*** – nominální proměnné – nerozlišují stupeň vlastnosti nebo číselnou hodnotu, vyjadřují pouze různost; jsou to především kvalitativní kategorie, čísla znamenají pouze kódy

        - ***Ordinal*** – ordinální proměnné – kategorie, jejichž číselné kódy znamenají uspořádání, stupeň vlastnosti, pořadí
        - ***Scale*** – číselné proměnné, jejich hodnoty můžeme aritmeticky zpracovávat: sčítat, odčítat, násobit, dělit, umocňovat
    - u textových proměnných volíme pouze ***Nominal*** a ***Ordinal***
    - volba je analytikovým rozhodnutím, lze ji v průběhu analýzy kdykoliv měnit
    - u některých procedur tato volba nemá žádný vliv, u některých však má zásadní restriktivní vliv (např. nelze propočítat průměry pro nominální proměnné) nebo, jako v modulu **Decision Trees**, přímo určuje metodu zpracování; týká se to jen vybraných procedur
- **Role** – role proměnné se týká jen některých modelovacích procedur;
    - role určuje postavení proměnné v modelu:
        - ***None*** – role není přiřazena
        - ***Input*** – proměnná je v modelu chápána jako vstupní, nezávislá, prediktor, faktor
        - ***Target*** – výstupní, cílová proměnná modelu, predikant, závislá proměnná
        - ***Both*** – proměnná vystupuje v modelu současně jako výstupní i vstupní
        - ***Partition*** – proměnná bude použita jako rozdělení souboru na tři části: *trénovací, testovací a validační*
        - ***Split*** – podpůrná role, zavedená vzhledem ke kompatibilitě s *IBM SPSS Modeler*; *nemá* funkci příkazu **Split File**
    - pro vybrané analytické procedury musí být tato role dobře nastavena, může však být v některých procedurách měněna i v jejich rámci

Tip: Pro roli *Partition* je vhodné generovat proměnnou založenou na náhodných číslech postupem v **Transform – Compute**.

Všechny parametry proměnných mohou být měněny během analýzy prostým přepisem volby v nabídce příslušných polí nebo volbou nabídky pole.

Sloupce v okně **Variable View** lze potvrzením v záhlaví a přesunem myší převést na jiné místo, a tak si uživatel může nastavit takové pořadí, které mu vyhovuje. Pro běžnou analýzu může být pohodlnější pořadí: *Name, Label, Values, Missing, Type, Measure*, … Změnu provedeme také v nabídce **View – Customize Variable View** nebo v nabídce **Edit – Options – Data – Customize Variable View**.

Proměnné mohou být přeuspořádány podle pořadí číselných a abecedních hodnot kteréhokoliv atributu ve **Variable View** vzestupně nebo sestupně – nabídku pro tento krok získáme kliknutím pravým tlačítkem myši na záhlaví sloupce.

Datasety

V analýze dat můžeme použít současně několik pracovních souborů, které jsou otevřeny a mezi nimiž můžeme libovolně přepínat a využívat je nezávisle na ostatních. Tyto soubory se nazývají *datasety.* Vznikají několika způsoby:

a) prázdný dataset dostaneme otevřením nového souboru
   File – New – Data
   používá se, po definování názvů sloupců, pro manuální vstup datové matice buď přímým pořízením, nebo kopírováním z vnějších zdrojů
b) existující soubor *.sav* otevřeme běžným způsobem v záložce ***File***
   File – Open – Data
c) soubory jiných typů (Excel, dBase, Lotus, Cognos, relační databáze pomocí ODBC) otevřeme běžným způsobem v záložce ***File – Open*** nebo ***File – Open Database*** a další možnosti v záložce: **File – Read Text Data**, **File – Read Cognos data**, **File – Read Triple-S data**, **File – Get R Space**; vzniklý dataset poté můžeme uložit ve formátu *.sav*
d) kopírujeme jiný dataset
   Data – Copy Dataset
   používá se jednak proto, aby původní soubor nebyl porušen a zůstal jako originál k dispozici, a jednak pro speciální úpravy, většinou pro zjednodušení u dílčích úloh analýzy:
   - redukce velkého počtu proměnných na soubor potřebný pro speciální dílčí úlohu
   - redukce dat na dílčí podsoubory, které jsou předmětem zájmu (např. podsoubory mužů a žen, podsoubory pacientů podle diagnóz, věkové kohorty apod.)
   - základ pro agregaci nebo transformaci, spojení souborů nebo rozšíření o proměnné jiného souboru
   - odvozené výsledky datového procesu, např. transponovaný soubor, výsledky imputace chybějících hodnot

Pro použití v dalším postupu dostávají datasety číslovaná jména tvaru *DataSet1*, *DataSet2*… (tyto názvy se používají také) v syntaktickém zadání. Původní název souboru se zobrazuje na horní liště spolu se jménem datasetu. Jméno datasetu můžeme změnit v nabídkovém okně v menu: **File – Rename Dataset –** ***nabídkové okno***

Transpozice

`Data – Transpose`

Transpozice je výměna řádků za sloupce a naopak. Původní proměnné se stávají případy a případy se stávají proměnnými.

Postup zadání:

1. Vybereme ty proměnné, které budou reprezentovat budoucí řádky.
2. Určíme proměnnou, která určí jména nových proměnných v transponované matici. Je-li to textová proměnná, budou její hodnoty použity jako názvy (s nutnou změnou, např. mezery jsou vyplněny podtržníkem), u číselných proměnných se přidává písmeno na začátek.
3. Po potvrzení ***OK*** se nový soubor objeví v novém datasetu.
4. Všechny případy se staly proměnnými, vybrané proměnné se staly případy/řádky. Názvy původních proměnných tvoří první textovou proměnnou CASE_LBL.

Ke stejnému nabídkovému oknu a postupu se dostaneme postupem:
`Data – Restructure ... – Transpose All Data – Finish` – nabídkové okno

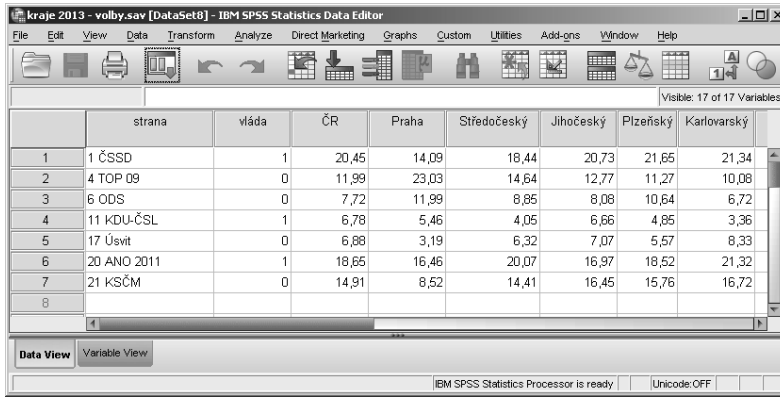
Tip: Chcete-li převést na proměnné jen některé řádky, pak nejprve zkopírujte dataset, vyberte v něm řádky pomocí výběru **Select Cases** v záložce **Data** nebo v menu **Edit – Clear** nebo pomocí tlačítka klávesnice ***Delete*** odstraňte příslušné řádky a poté aplikujte postup transpozice tabulky.

Příklad 1.1: *Strany vs. kraje*

Změna řádků na sloupce znamená, že budeme zpracovávat kraje jako případy a výsledky stran jako proměnné.

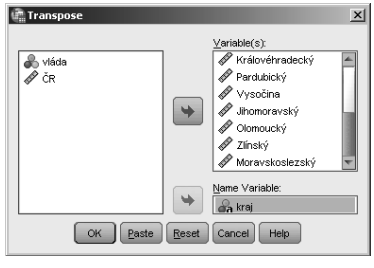
Postup je vidět v obrázcích 1.9, 1.10 a 1.11. Po **`Data – Transpose`** zvolíme v okně proměnné, které se transponují do případů. Budou to všechny kraje, ale vynecháme sloupec ČR, protože ten mezi případy (tj. kraje) nepatří. (Pokud bychom jej také transponovali, museli bychom jej posléze z tabulky odstranit.)

V souboru „*kraje 2013 - volby.sav*" jsou v řádcích reprezentovány parlamentní strany a ve sloupcích výsledky voleb v krajích.

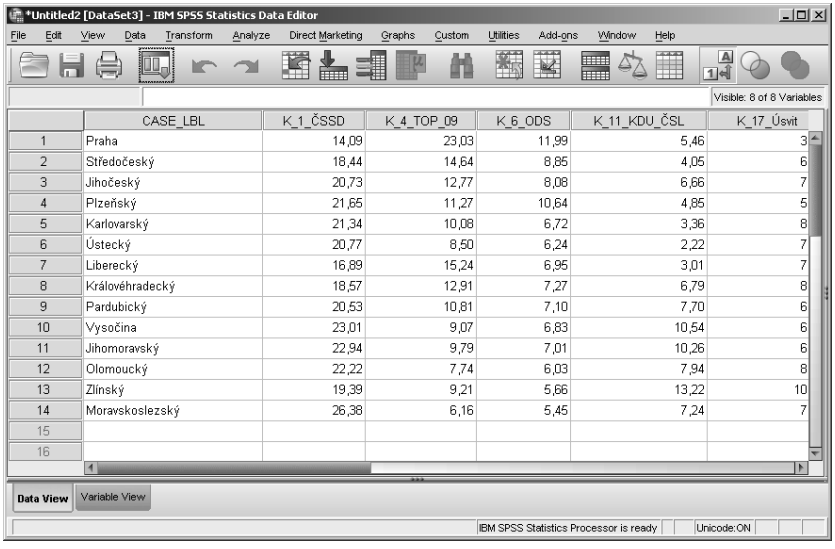


	strana	vláda	ČR	Praha	Středočeský	Jihočeský	Plzeňský	Karlovarský
1	1 ČSSD	1	20,45	14,09	18,44	20,73	21,65	21,34
2	4 TOP 09	0	11,99	23,03	14,64	12,77	11,27	10,08
3	6 ODS	0	7,72	11,99	8,85	8,08	10,64	6,72
4	11 KDU-ČSL	1	6,78	5,46	4,05	6,66	4,85	3,36
5	17 Úsvit	0	6,88	3,19	6,32	7,07	5,57	8,33
6	20 ANO 2011	1	18,65	16,46	20,07	16,97	18,52	21,32
7	21 KSČM	0	14,91	8,52	14,41	16,45	15,76	16,72
8								

Obrázek 1.9 Dataset *„kraje 2013 - volby.sav"*



Obrázek 1.10 Zadání transpozice pro zpracování souboru krajů



	CASE_LBL	K_1_ČSSD	K_4_TOP_09	K_6_ODS	K_11_KDU_ČSL	K_17_Úsvit
1	Praha	14,09	23,03	11,99	5,46	3
2	Středočeský	18,44	14,64	8,85	4,05	6
3	Jihočeský	20,73	12,77	8,08	6,66	7
4	Plzeňský	21,65	11,27	10,64	4,85	5
5	Karlovarský	21,34	10,08	6,72	3,36	8
6	Ústecký	20,77	8,50	6,24	2,22	7
7	Liberecký	16,89	15,24	6,95	3,01	7
8	Královéhradecký	18,57	12,91	7,27	6,79	8
9	Pardubický	20,53	10,81	7,10	7,70	6
10	Vysočina	23,01	9,07	6,83	10,54	6
11	Jihomoravský	22,94	9,79	7,01	10,26	6
12	Olomoucký	22,22	7,74	6,03	7,94	8
13	Zlínský	19,39	9,21	5,66	13,22	10
14	Moravskoslezský	26,38	6,16	5,45	7,24	7
15						
16						

Obrázek 1.11 Výsledný soubor transpozice: kraje jsou případy, volební výsledky jsou sloupce, názvy řádků v CASE_LBL jsou převzaty z názvů proměnných původního souboru

Typickou úlohou pro transponování matice dat jsou expertní hodnocení jednotlivých položek, např. produktů, strategií, budoucích scénářů apod. Dvojí pohled vede buď na komparaci expertů (řádek reprezentuje hodnocení položky experty), nebo na komparaci hodnocených položek (řádek reprezentuje hodnocení položek expertem).

Jinou typickou úlohou je záměna respondentů řádků původního souboru dat a položek baterie dotazníkových otázek (sloupce) pro aplikaci Q-metodologie faktorové analýzy, ve které se vychází z korelační matice mezi respondenty, tedy matice spočtené po transpozici.

Restrukturace

`Data – Restructure...`

Postupy restrukturace přemění záznamy tak, aby nově vytvořená forma souboru odpovídala vhodným způsobem dané analytické úloze. Patří sem dva zásadní způsoby přeorganizování dat:

A. převedení vybraných proměnných na případy

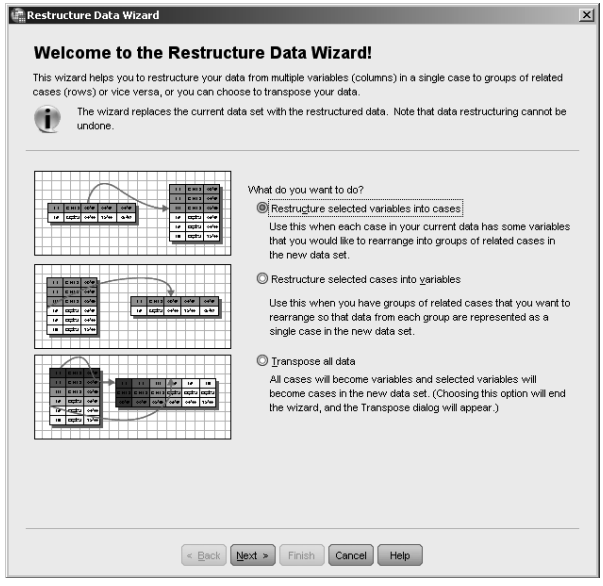
B. převedení vybraných případů na proměnné

K restrukturačním postupům patří také záměna řádků a sloupců matice dat (viz část *Transpozice* v této kapitole).

Nově vytvořený soubor se při restrukturaci zobrazí v novém datasetu, takže původní soubor zůstane zachován.

Restrukturační postupy jsou voleny v nabídce:

`Data – Restructure ...` – ***výběr způsobu restrukturace*** – další postup podle průvodce a jeho voleb.



Obrázek 1.12 Vstupní okno průvodce zadáváním restrukturace

Volba **`Data – Restructure ...`** – ***Transpose all data*** je realizována stejným způsobem, který byl popsán v části *Transpozice* (viz výše v této kapitole).

A. volba restrukturace: z vybraných proměnných vzniknou případy
Restructure selected variables into case

A1) jedna skupina proměnných
How many variable groups: One

V řádcích matice může být několik pozic, jejichž záznamy mají takový charakter, že je chceme sloučit do jednoho sloupce (a zmnožit tak řádky), abychom mohli takovou proměnnou zpracovat souhrnně. K úlohám souhrnného zpracování několika proměnných současně máme k dispozici buď definici násobných odpovědí a k nim příslušných tabelačních metod (*multiple response*, viz kapitola 2), nebo metodu restrukturace, která ze skupiny proměnných vytváří proměnnou jednu převodem na případy.

Převod *skupinky proměnných* (*variable group*) na *skupinku případů* (*case group*) spočívá v tom, že každý případ opakujeme v novém souboru tolikrát, kolik proměnných je ve skupině, zavedeme jednu novou proměnnou (*Target variable*) a v ní se postupně uvedou hodnoty zvolených proměnných z převáděné skupiny v pořadí, jak jsou zapsány v souboru. Řádkové hodnoty se přesunou do jednoho sloupce k nově vzniklým řádkům („zmnoží se případ"). Tím se z několika sloupců stane jeden a každý případ se zmnoží na skupinku případů tolikrát, kolik proměnných převádíme. Vedle sloupce cílové proměnné zavádíme také indexovou proměnnou, která určuje, ze které proměnné je hodnota přejata.

V novém souboru mohou zůstat také ostatní proměnné (mimo restrukturovanou proměnou), a to buď všechny, nebo jen vybrané.

Identifikace vzniklých skupinek nových případů je uložena, pokud chceme, v nové *identifikační proměnné* s volitelným jménem a s hodnotami buď z nějaké existující proměnné, nebo s pořadovým číslem případu v původní.

V řadě aplikací je také vhodné zjišťovat velikost nově vzniklé skupinky případů (při volbě redukce souboru o systémově vynechané hodnoty či prázdná pole).

Postup zadání:

1. Určíme skupinku proměnných a jméno cílové proměnné v rámečku ***Variables to be Transposed*** a případně pomocí šipek změníme jejich pořadí (pořadí proměnných určuje pořadí případů v odvozeném souboru).
2. Určíme jednoznačnou identifikaci odvozené skupinky případů v rámečku ***Case Group Identification***, a to buď číslem pořadí případu (***Use case number***), nebo zvolenou proměnnou (***Use selected variable***); u identifikační proměnné zvolíme jméno a případně přiřadíme název. Volba ***None*** znamená, že identifikaci neprovádíme.
3. Zvolíme *fixní proměnné*, které v souboru zůstanou a jejichž hodnoty se budou opakovat pro každý ze skupinky případů, pokud chceme ponechat jen některé.
4. V okně ***Variables to Cases: Create Index Variables*** rozhodneme, zda zavedeme *indexovou proměnnou*, která určuje, z jaké proměnné případ vznikl.
5. V okně ***Variables to Cases: Create One Index Variable*** určíme *indexovou proměnnou*. V rámečku ***What kind of index variable?*** rozhodneme, zda použijeme pořadové číslo proměnné ve skupince (***Sequential numbers***) nebo jména proměnných (**Variable names**). Určíme jméno indexové proměnné.
6. V okně ***Variables to Cases: Options***, v rámečku ***Handling of Variables not Selected***, volíme akci pro nerestrukturované proměnné původního souboru:
   a) ***Drop variable(s) from the new file*** – nepřevádíme je,
   b) ***Keep and treat as fixed variables*** – nepřeváděné proměnné zůstanou také v odvozeném souboru jednotlivě (a pro skupinu případů k jednomu identifikačnímu číslu se budou opakovat).
7. Ve stejném okně v rámečku ***System Missing or Blank Values in all Transposed Variables*** rozhodneme, zda budou zavedeny i případy se systémově vynechanými hodnotami pro numerické proměnné a prázdnými poli pro textové proměnné (***Create a case in the new file***), nebo zda budou vynechány (***Discard the data***).
8. V rámečku **Case Count Variable** volíme zařazení proměnné „velikost skupinky vzniklých případů".
9. Provedeme akci nebo přejdeme do okna ***Finish***, ve kterém buď provedeme restrukturaci (***Restructure the data now***), nebo zapíšeme pokyn restrukturace v syntaxovém okně (***Paste …***).

Tip: Protože původní soubor bude přepsán novým restrukturovaným souborem, je vhodné si vytvořit nejprve duplikát původního souboru pomocí **Copy Dataset** a provádět restrukturaci na něm.

Využití restrukturace v běžné praxi:

Příklad 1.2: *Výzkum automobilového trhu*

Otázka: „Jaké značky automobilu vlastníte ve Vaší rodině?"

U každého respondenta (řádku datové matice) zaznamenáme kódy značek aut pro všechna vozidla, které rodina vlastní, v pořadí, jak je uvedl respondent, postupně do sloupců *Auto1* až *Auto5*. Nevyplněné sloupce budou mít určeny systémově vynechané hodnoty.

Analytická otázka zní: Jak charakterizovat majitele určité značky a komparovat značky mezi sebou? Jednotlivé značky jsou rozhozeny v pěti sloupcích, což činí analýzu komplikovanou.

Při průchodu průvodcem pojmenujeme cílovou proměnnou jako „*značka*", určíme indexování názvem proměnné (pokud jej vůbec vyžádáme), určíme také proměnnou „*počet*", která provádí počet automobilů v rodině, a vyloučíme případy, které by odpovídaly systémově vynechaným hodnotám u původních proměnných. Všechny ostatní původní proměnné necháme fixní, tj. u všech vozidel jedné rodiny se údaje budou opakovat. Rodiny, které žádné auto nevlastní, v novém souboru vynecháme.

Restrukturace souboru převede skupinu proměnných na případy a tak dostaneme soubor vlastněných aut všech respondentů. Počet řádků je roven počtu všech automobilů zaznamenaných v původním souboru.

Vzniklé přeorganizování dává možnost třídit data podle značky vozidla, komparovat procenta, průměry a jiné deskriptivní charakteristiky pro jednotlivé značky, agregovat podle výrobce apod. Model je běžně používán ve výzkumu hodnocení produktů spotřebiteli, při výzkumu nákupního chování a při podobných úlohách výzkumu trhu.

Příklad 1.3: *Výzkum zájmů ve volném čase*

Otázka v e-výzkumu: „Jaké zájmové činnosti a koníčky provádíte ve volném čase?" „Co je pro Vás nejčastější, co na druhém místě, třetím…"

Záznam odpovědí je pro každého respondenta zapsán v jeho řádku matice dat v proměnných *Volnyčas1* až *Volnyčas10*. Cílová proměnná bude nazvána např. *Aktivita*. Postup je zcela stejný jako v příkladu 1.2, jen místo indexování názvem proměnné zvolíme indexování pořadím proměnné ve skupině. Tím získáme dodatečnou informaci o pořadí (číselná proměnná) pro další zpracování. Nový soubor je souborem všech volnočasových aktivit zaznamenaných v souboru.

Obdobné záznamy mohou zachytit návštěvy koncertů nebo výstav, výčet navštívených hradů a zámků za poslední rok atd.

Příklad 1.4: *Záznam věku dětí při výzkumu rodiny*

Řádek je záznamem pro rodinu a věky dětí jsou zapsány ve sloupcích *Dítě1* až *Dítě10*.

Indexování provedeme podle pořadí proměnné (tj. podle pořadí věku dítěte). Cílová proměnná je číselná a lze ji zpracovávat nejen nominálně (kód pořadí), ale i numericky (hodnota pořadí).

Postupy restrukturace od proměnných k případům se používají často u tzv. otevřených otázek v dotazníku (vyjmenujte všechny čisticí prostředky, které byly zakoupeny poslední měsíc; vyjmenujte politiky, kteří Vám jsou sympatičtí) a jiných, předem a nestrukturovaných záznamů (soupis vybavení bytu, seznam trestních činů, seznam chorob v anamnéze…).

A2) Více skupin proměnných

How many variable groups: More than one

Při restrukturaci proměnných na případy lze volit více skupin. Postup se liší jen v tom, že v postupném návodu přepneme v okně ***Variables to Cases: Number of Groups*** předvolbu na ***More than one***, určíme počet skupin a v dalším okně je postupně definujeme v okně ***Target Variable***.

Počet proměnných v každé skupině musí být stejný.

Postup poskytne nové cílové proměnné v zadaném počtu. Z jednoho původního řádku vznikne tolik nových případů, kolik je společný počet proměnných ve skupinách. K prvnímu odvozenému případu budou přiřazeny hodnoty z prvních proměnných v každé skupině, k druhému to budou hodnoty z druhých proměnných skupiny atd.

Příklad 1.2 (pokračování):

Výzkum automobilového trhu – otázka „Jaké značky automobilu vlastníte ve Vaší rodině?" je doplněna otázkami: „Hodnoťte postupně uvedená auta na stupnici 1 – 7: komfort řízení", „Hodnoťte postupně uvedená auta na stupnici 1 – 7: kvalita motoru", „Hodnoťte postupně uvedená auta na stupnici 1–7: spotřeba", „Hodnoťte postupně uvedená auta na stupnici 1 – 7: vzhled" a „Koupíte si tuto značku znovu?"

Ke každému vozidlu respondenta připadá pět dalších otázek, ale v daném tvaru záznamu se data zpracovávat nedají. Proto záznamy přeorganizujeme restrukturací:

a) definujeme šest skupin proměnných k převodu na případy – seznam automobilů (Auto1 až Auto5), hodnoceni komfortu (Komfort1 až Komfort5), hodnocení kvality (Kvalita1 až Kvalita5), (Spotřeba1 až Spotřeba 5), (Vzhled1 až Vzhled 5), (Opaknakup1 až Opaknakup5)
b) postupujeme dále zcela stejně podle návodných oken jako u jedné skupiny
c) výsledkem jsou řádky nového souboru, ve kterém budou vždy uvedeny postupně v určených sloupcích: značka automobilu, hodnocení komfortu řízení, kvality motoru, spotřeby, vzhledu a ochota koupit značku znovu.

Tím dostaneme vedle sebe korespondující vlastnosti pro každé vozidlo a dobře zpracovatelný soubor.

Příklad 1.3 (pokračování):

Výzkum zájmů ve volném čase – otázky „Jaké zájmové činnosti a koníčky provádíte ve volném čase? – Co je pro Vás nejčastější, co na druhém místě, třetím…?" jsou doplněny otázkami pro každou aktivitu: „Kolik času věnujete těmto aktivitám" a „Děláte tyto činnosti sám/sama, nebo s některým členem rodiny nebo s přítelem/přítelkyní?

Podobně jako v příkladu 1.2 (pokračování) převedeme tři skupiny proměnných na případy a můžeme v analýze vztáhnout jednotlivé typy aktivit k intenzitě realizace, k roli typu aktivity v rodině či k měření souladu zájmů partnerů.

1.4 (pokračování): Věk dětí

U dětí můžeme dále zaznamenat postupně jejich vlastnosti: pohlaví, váhu, výšku, typ školy, datum narození, průměrnou známku ve škole…

Restrukturací dostaneme řádky, které odpovídají dětem. Záznamy jsou organizovány jako vlastnosti dětí a jsou snadno zpracovatelné (včetně korelací a komparací). Indexování pořadím dítěte také otevře analýzu dětí podle pořadí narození.

B) Převedení vybraných případů na proměnné

Restructure selected variables into cases

Převádění případů na proměnné se provádí v případech, kdy několik řádků obsahuje informaci o jedné statistické jednotce, nebo je-li několik případů spojeno klasifikačními proměnnými. Naším cílem je převést tuto informaci do jednoho řádku, tj. vytvořit nový soubor, v němž spojená informace tvoří jeden řádek. Informace se převádějí z původních řádků (případů výchozího souboru) do proměnných (sloupců).

Pro jednu statistickou jednotku zjišťujeme hodnoty proměnné opakovaně v různých situacích nebo okamžicích a v datové matici jsou data pro každou jednotku zaznamenána postupně tak, že každý záznam má svůj vlastní řádek. Naším cílem je převést tuto informaci do jednoho řádku, tj. vytvořit nový soubor, v němž statistická jednotka tvoří jeden řádek matice a hodnoty proměnných z původních řádků se převádějí do sloupců.

Typickými jsou např. opakovaná měření v čase, panelová šetření při výzkumu trhu, transakční databáze jako záznamy o nákupech u jedné věrnostní karty, záznamy o výběrech a vkladech bankovního účtu, návštěvy pacienta u lékaře, denní pozorování počasí v souboru různých stanic apod.

Modelovým příkladem je opakované měření fyziologických veličin u pacientů v definovaných časových okamžicích nebo různými lékařskými metodami. U pacienta je zaznamenán krevní tlak před podáním léčebného přípravku, hodinu po podání a 3 hodiny po podání. Pro každého pacienta jsou data zaznamenána v matici třemi řádky pro každé měření zvláště se třemi proměnnými: identifikace pacienta, doba měření a hodnota tlaku. Matici ale z důvodů statistického zpracování převedeme na tvar, v němž všechna tři měření tlaku pro každého pacienta budou umístěna v jednom řádku vedle sebe ve třech sloupcích odpovídajících jednotlivým dobám měření. Situaci ilustruje následující schéma.

PACIENT	DOBA	TLAK
ID1	před	150
ID1	po 1 h	130
ID1	po 3 h	120
ID2	před	140
ID2	po 1 h	130
ID2	po 3 h	115

PACIENT	TLAK před	TLAK po 1 h	TLAK po 3 h
ID1	150	130	120
ID2	140	130	115

Proces restrukturace vyžaduje identifikaci statistické jednotky, tedy nositele informace, a přiřazení informace společným situacím. K tomu slouží proměnné *index* a *identifikátor*.

Identifikátor **(Identifier Variable(s))** – je proměnná definující skupinu řádků, které patří ke stejné statistické jednotce. Často to bývá ID subjektu. Statistická jednotka může být určena i kombinací více proměnných. Každá hodnota nebo kombinace hodnot identifikátoru bude tvořit jeden řádek v restrukturalizované matici. (Za jednotku považujeme např. jeden měsíc a ten je určen kombinací roku a měsíce v roce.)

Index **(Index Variable(s))** – je proměnná (resp. proměnné) určující řádky u jednotlivých statických jednotek, které se týkají jedné situace měření. Každá hodnota nebo kombinace hodnot indexu bude v restrukturalizované matici tvořit jeden sloupec. (U opakovaného měření u pacientů index udává, zda byla hodnota naměřena před léčbou nebo po léčbě. Index může být založen i na více proměnných; u pacientů je kromě času měření další indexovou proměnnou rozlišena metoda měření.)

Všechny řádky s jednou hodnotou identifikátoru v originální matici jsou v restrukturalizované matici přeuspořádány do jednoho řádku. Z každé restrukturované proměnné originální matice vznikne sada proměnných, pro každou hodnotu indexu jedna. Každá indexová hodnota tak vytvoří jeden sloupec, který bude pojmenován podle hodnoty indexové proměnné. Pokud pro nějakou hodnotu identifikátoru chybí jedna nebo více hodnot indexu, bude v nové matici pro tuto kombinaci uvedena systémově chybějící hodnota. (U některých pacientů nejsou známy hodnoty před léčbou nebo nebyly měřeny všemi způsoby.)

Protože identifikátor určuje v nové matici jeden řádek a index určuje jeden sloupec, identifikují oba v kombinaci jedno pole nové datové matice. Tím je údaj jednoznačně určen ve struktuře nového souboru.

Postup zadání:

1. V okně ***Cases to Variables: Select Variables*** určíme identifikaci statistické jednotky (*identifikátor*) jednou proměnnou nebo jednoznačnou kombinací více proměnných v okně ***Identifier Variable(s)***.
2. Určíme identifikaci údajů patřících k sobě (i*ndex*) v okně ***Index Variable(s)***.
3. V okně ***Cases to Variables: Sorting data*** soubor setřídíme podle identifikátorů a indexů: zpravidla ponecháme předvolbu ***Yes***; je-li soubor již setříděn, můžeme třídění vynechat, volba ***No*** je ale vhodná pro velké soubory, kdy je přetřídění dat náročné na čas.
4. V okně ***Cases to Variables: Options*** určíme:
   a) v rámečku ***Order of New Variable Groups*** pořadí proměnných a jejich skupiny v nové datové matici:
      i) podle původního pořadí ve výchozích datech (volba ***Group by original variable***) – vytvořené proměnné jsou seskupeny vždy podle původní proměnné, ze které vznikly
      ii) podle hodnot indexu (volba ***Group by index***) – odvozené proměnné jsou seřazeny ve skupinách ke každé hodnotě indexu
   b) v rámečku ***Count the number of cases*** volitelnou proměnnou udávající počet řádků původní matice, které se restrukturalizovaly do jednoho řádku nové matice. Je to počet hodnot indexu u identifikátoru.
   c) v rámečku ***Create indicator variables*** volitelné indikátory, které udávají (hodnotami 0 a 1), zda se v originální matici vyskytuje určitá kombinace indexu a identifikátoru.
5. Restrukturalizaci buď ihned provedeme tlačítkem ***Finish***, nebo přejdeme do okna ***Finish*** pro uložení pokynu do syntaktického okna.

Poznámka: Proměnné, které nebyly v postupu použity jako indikátory a mají shodné hodnoty indexů pro každou skupinu proměnných, se reprodukují touto hodnotou do odvozeného řádku. Všechny proměnné, které u některé statistické jednotky mají různé hodnoty, vstoupí do restrukturace podle indexu.

Tip: Chcete-li si zachovat původní soubor, restrukturujte data na jeho kopii ve formátu *.sav* nebo zkopírujte pracovní soubor do nového datasetu.

Příklad 1.5: *Opakovaná měření*

Soubor („*měření_hmotnosti.sav*“) obsahuje data z opakovaných měření pacientů dietní studie. Každé měření je zaznamenáno ve vlastním řádku. Měřena byla hmotnost („*hmotnost*“) a obsah triglyceridů („*tg*“). Měření se provedlo ve třech okamžicích: před dietou, v průběhu diety a po dietě. Okamžik měření zachycuje proměnná „*čas*“. Soubor obsahuje i statická data o pacientech, která se u pacienta nemění, např. „*pohlaví*“ a „*vzdělání*“.

měření_hmotnosti.sav [DataSet5] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Custom Utilities Add-ons Window Help

Visible: 9 of 9 Variables

	id	pohlaví	věk	vzdělání	onemocnění	diety	čas	hmotnost	tg
1	1	muž	do 30 let	základní	skupina 1	B	před začát...	71,0	163,80
2	1	muž	do 30 let	základní	skupina 1	B	po ukonče...	63,8	170,06
3	1	muž	do 30 let	základní	skupina 1	B	po ukonče...	62,0	171,21
4	2	žena	do 30 let	základní	skupina 2	B	před začát...	68,1	158,87
5	2	žena	do 30 let	základní	skupina 2	B	po ukonče...	65,7	161,04
6	2	žena	do 30 let	základní	skupina 2	B	po ukonče...	60,1	166,26
7	3	žena	61 a více	základní	skupina 2	B	před začát...	69,3	157,21
8	3	žena	61 a více	základní	skupina 2	B	po ukonče...	62,1	164,59
9	3	žena	61 a více	základní	skupina 2	B	po ukonče...	61,0	164,72
10	4	žena	do 30 let	vyskoškolské	skupina 2	B	před začát...	70,2	156,40
11	4	žena	do 30 let	vyskoškolské	skupina 2	B	po ukonče...	67,2	158,65
12	4	žena	do 30 let	vyskoškolské	skupina 2	B	po ukonče...	62,8	164,38
13	5	žena	do 30 let	vyskoškolské	skupina 2	B	před začát...	70,6	157,43
14	5	žena	do 30 let	vyskoškolské	skupina 2	B	po ukonče...	66,3	159,86
15	5	žena	do 30 let	vyskoškolské	skupina 2	B	po ukonče...	64,3	161,25
16	6	žena	do 30 let	středoškolské	skupina 1	A	před začát...	72,0	155,76

Data View Variable View

IBM SPSS Statistics Processor is ready Unicode:ON

Obrázek 1.13 Data o hmotnosti a obsahu triglyceridů v původním záznamu jednotlivých měření

Pro analýzu dat a komparaci změn restrukturujeme soubor tak, aby každý pacient měl všechny své záznamy v jednom řádku. Záznamy o hmotnosti a o triglyceridech budou umístěny ve sloupcích odvozené datové matice podle toho, zda byla měření realizována před, v průběhu či po skončení diety. K tomu budou přidány sloupce s osobními charakteristikami, které se ve výchozím souboru v záznamech pacientů nemění. Nový formát je vhodný pro různé komparační a korelační výpočty a analýzy a pro aplikaci modelů opakovaných měření ANOVA.

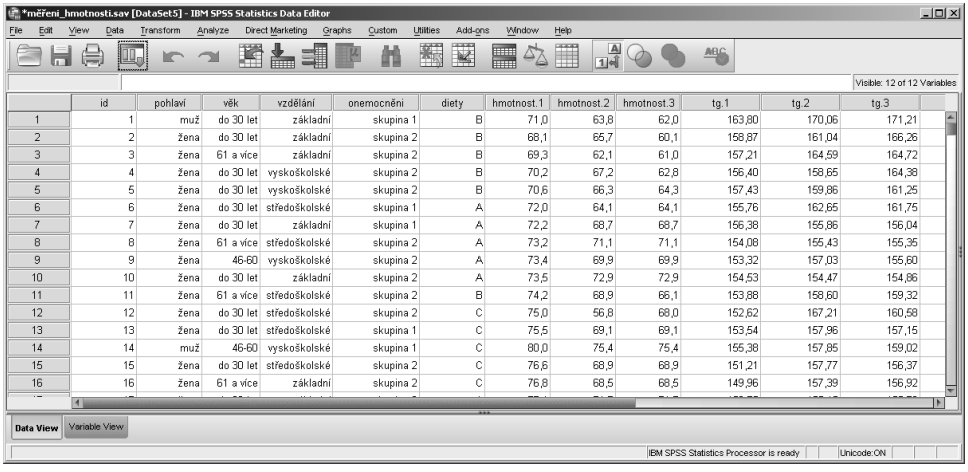
Zvolíme **`Data – Restructure – Restructure selected cases into variables`** a pokračujeme postupně nabídkovými okny:

1. Identifikátorem bude proměnná „*id*“ – převedeme ji v nabídce ***Cases to Variables: Select Variables*** do okna ***Identifier Variable(s)***.
2. Indexem bude proměnná „*čas*“, kterou převedeme do okna ***Index Variable(s)***
3. V okně ***Cases to Variables: Sorting data*** ponecháme předvolbu ***Yes***.
4. V okně ***Cases to Variables: Options*** určíme: v rámečku ***Order of New Variable Groups*** pořadí proměnných a jejich skupiny v nové datové matici, podle původního pořadí ve vý-

chozích datech (volba ***Group by original variable***); ostatní volby vynecháme, protože ani počet případů ve skupině, ani indikační proměnné nás v této úloze nezajímají.

5. Potvrdíme tlačítko ***Finish*** a akce proběhne.

Informace o každém pacientovi je ve vzniklém souboru na jednom řádku. Z proměnných „*hmotnost*“ a „*tg*“ vznikly trojice proměnných: měření před, v průběhu a po dietě. Jména nových proměnných obsahují číslo odpovídající kódu času měření. Statické proměnné jsou v souboru jen jednou, program zjistil, že jejich hodnoty jsou u každého pacienta konstantní, a nevytvořil z nich sady proměnných.



	id	pohlaví	věk	vzdělání	onemocnění	diety	hmotnost.1	hmotnost.2	hmotnost.3	tg.1	tg.2	tg.3
1	1	muž	do 30 let	základní	skupina 1	B	71,0	63,8	62,0	163,80	170,06	171,21
2	2	žena	do 30 let	základní	skupina 2	B	68,1	65,7	60,1	158,87	161,04	166,26
3	3	žena	61 a více	základní	skupina 2	B	69,3	62,1	61,0	157,21	164,59	164,72
4	4	žena	do 30 let	vyskoškolské	skupina 2	B	70,2	67,2	62,8	156,40	158,65	164,38
5	5	žena	do 30 let	vyskoškolské	skupina 2	B	70,6	66,3	64,3	157,43	159,86	161,25
6	6	žena	do 30 let	středoškolské	skupina 1	A	72,0	64,1	64,1	155,76	162,65	161,75
7	7	žena	do 30 let	základní	skupina 1	A	72,2	68,7	68,7	156,38	155,86	156,04
8	8	žena	61 a více	středoškolské	skupina 2	A	73,2	71,1	71,1	154,08	155,43	155,35
9	9	žena	46-60	vyskoškolské	skupina 2	A	73,4	69,9	69,9	153,32	157,03	155,60
10	10	žena	do 30 let	základní	skupina 2	A	73,5	72,9	72,9	154,53	154,47	154,86
11	11	žena	61 a více	středoškolské	skupina 2	B	74,2	68,9	66,1	153,88	158,60	159,32
12	12	žena	do 30 let	středoškolské	skupina 2	C	75,0	56,8	68,0	152,62	167,21	160,58
13	13	žena	do 30 let	středoškolské	skupina 1	C	75,5	69,1	69,1	153,54	157,96	157,15
14	14	muž	46-60	vyskoškolské	skupina 1	C	80,0	75,4	75,4	155,38	157,85	159,02
15	15	žena	do 30 let	středoškolské	skupina 2	C	76,6	68,9	68,9	151,21	157,77	156,37
16	16	žena	61 a více	základní	skupina 2	C	76,8	68,5	68,5	149,96	157,39	156,92

Obrázek 1.14 Data o hmotnosti a obsahu triglyceridů po restrukturaci podle jednotlivců

Příklad 1.6: *Transakční data – záznamy o nákupech*

Originální soubor („*Transakce.sav*“) obsahuje transakční data o nákupech z prodejny supermarketu. Řádek v matici tvoří ID nákupu, kategorie položky zboží v nákupu a proměnné jsou „*počet*“ (počet položek) a „*částka*“ (částka za položku). Pro zhodnocení souhrnných nákupů spojíme jednotlivé informace do jednoho společného řádku. Restrukturovaná forma je nutná pro analýzu nákupního košíku a výpočty podílových a součtových ukazatelů.

Statistickou jednotkou je původně jeden nákup a v jeho rámci jsou uvedeny kategorie položek a jejich parametry v nákupu. Na rozdíl od předcházejícího příkladu 1.5 není struktura nákupu pevná, ale různé nákupy se skládají z různých kombinací položek. Nové statistické jednotky po přestrukturování budou zákazníci a položky se přesunou do sloupců.

1. Zvolíme **`Data – Restructure – Restructure selected cases into variables`** a pokračujeme tlačítkem **Next**.
2. Určíme strukturu nové datové matice.
   a) Řádek nové matice bude tvořit nákup, proto do pole **Identifier Variable(s)** vložíme proměnnou „*ID*“.
   b) Sloupce nové matice budou kategorie položek, do pole **Index variables** vložíme proto proměnnou „*položka*“.

3. V dalším okně ***Cases to Variables: Sorting data*** ponecháme předvolbu **Yes**.
4. V okně ***Cases to Variables: Options*** určíme rozmístění nových proměnných a vytvoříme další proměnné.
   a) V původní matici jsou proměnné *„počet“* a *„částka“*; z každé z nich vznikne sada proměnných pro jednotlivé položky. Pro lepší orientaci necháme v nové matici pohromadě skupinu proměnných vzniklou z originální proměnné. V oblasti **Order of New Variable Groups** proto zvolíme **Group by Original Variable**.
   b) Abychom věděli, kolik druhů položek je v nákupu, vytvoříme v oblasti **Case Count Variable** proměnnou *„počet“*. Proměnná zachytí počet hodnot indexu (*položka*) v rámci identifikátoru (*ID*).
   c) Strukturu nákupu podle výskytu kategorie položky zachytí indikátorové proměnné. Začátek jména proměnných *kat* určíme v oblasti **Indicator Variable**.
5. Restrukturaci provedeme nyní tlačítkem ***Finish***.

V novém souboru je jeden nákup na jednom řádku. Proměnná *„počet“* udává počet kategorií položek v nákupu. Sada proměnných s *„kat1“* až *„kat10“* jsou indikátory výskytu kategorií zboží v nákupu. Čísla odpovídají kódování originální proměnné *„položka“*. Následuje restrukturovaná informace o počtu kusů v jednotlivých kategoriích a poté zaplacené částky. Pokud se v daném nákupu položka nevyskytovala, je v nové matici systémově vynechávaná hodnota (tu můžeme pro praktické cíle převést rekódováním na nulu (viz kap. 3).

Tip: Chcete-li mít v nových proměnných textový název položky, a ne kód, vytvořte v originálním souboru textovou proměnou z názvů kategorií (funkce VALUELABEL) a restrukturalizaci proveďte podle textové proměnné. Původní proměnnou je vhodné smazat, aby zbytečně nekomplikovala nový soubor. Pořadí nových proměnných bude podle abecedního pořadí textové proměnné v indexu.

Spojování souborů

`Data – Merge Files – Add Cases`

`Data – Merge Files – Add Variables`

Spojování pracovních souborů představuje rozšíření aktivního pracovního souboru a spojování informací z důvodu jejich společného zpracování. To lze provádět vcelku jednoduše – byť zdlouhavě – manuálně cestou přenosu jiného datasetu, *.sav* souboru nebo listu v souboru MS Excel do aktivního souboru/datasetu pomocí Copy/Paste, avšak rychlé softwarové řešení je mnohem vhodnější – otevírá více možností a eliminuje rizika chyb. Spojování souborů představuje tři možné operace:

a) připojení souboru nových případů s (částečně) stejnými proměnnými k pracovnímu souboru,
b) připojení souboru se stejnými případy a jinými proměnnými,
c) připojení dodatečné informace (např. agregované) k případům podle klíčových proměnných z jiného souboru.

Připojujeme tedy buď případy (ad a)), nebo proměnné (ad b) a c)), řádky nebo sloupce. Tyto operace můžeme nazvat rozšíření *„do délky“* a *„do šíře“* datové tabulky.

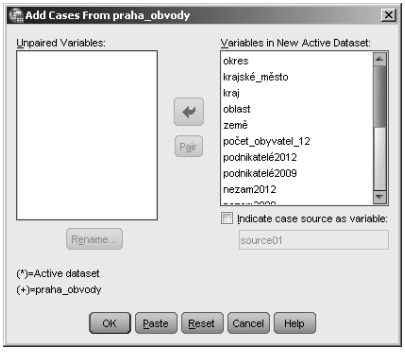
A) Připojení souboru nových případů s (částečně) stejnými proměnnými k pracovnímu souboru

`Data – Merge Files – Add Cases`

Připojení datových souborů *„do délky"*, tj. přidávání případů z vnějšího souboru k aktivnímu pracovnímu datasetu, předpokládá navázání proměnných obou souborů. Proto ty proměnné, které budou tvořit v souborech napojené sloupce, musí být pojmenovány stejně.

Postup zadání:

1. určení výchozího datasetu (jeho aktivice), ke kterému bude připojen jiný soubor
2. určení souboru, který bude připojen v okně ***Add Cases to „název aktivního datasetu"***
3. potvrdíme jeden z otevřených datasetů v konkrétní analytické situaci (jsou všechny vyjmenovány v okně) nebo
4. lokalizujeme a otevřeme externí soubor **IBM SPSS Statistics** (připojení souborů jiných typů než *.sav* musí předcházet jejich otevření ve formátu *sav*)
5. v dalším okně ***Add Cases From*** *„název připojovaného datasetu"* jsou soupisy
6. všech společných proměnných, které budou na sebe navazovat (***Variables in New Dataset***), a
7. proměnných, jejichž jména jsou jen v jednom souboru (***Unpaired Variables)***; proměnné aktivního souboru, které v napojovaném chybí, jsou označeny (*) a ty, které chybí v aktivním a jsou zahrnuty v napojovaném, se značí (+)
8. označením nepárované proměnné a tlačítkem ***Rename*** lze změnit jméno proměnné (a tak zajistit párování) pro případ, že stejná proměnná je pojmenována v obou souborech různě
9. zaškrtnutím a určením názvu zavedeme do odvozeného souboru dodatečnou proměnnou indikující zdroj údaje v řádku (0 = aktivní soubor, 1 = připojovaný soubor)



Obrázek 1.15 K souboru mimopražských okresů je připojován soubor pražských obvodů

Nepárované proměnné se také přenášejí do spojeného souboru. Vytvoří sloupec pod původním jménem, avšak hodnoty jsou uvedeny jen u případů souboru, ze kterého proměnná pochází, u druhého souboru jsou zavedeny systémově vynechávané hodnoty.

Příklad 1.7: *Spojení informací o Praze a mimopražských okresech*

Spojovat budeme soubory *„obvody Prahy 2012 – charakteristiky.sav"* a *„okresy mimo Prahu 2012 – charakteristiky.sav"*. Zvolíme první soubor jako aktivní a druhý připojíme. K tomu ale musí být proměnné, které vyjadřují stejnou informaci, pojmenovány stejně. V tomto případě je pojmenování v pořádku a můžeme soubor vytvořit a případně uložit.

Tip: Při přípravě souboru, který má být později napojen, zavádějte už ve fázi přípravy jména proměnných tak, aby bylo zajištěno párování, tedy stejně jako u souboru, ke kterému budete napojovat.

Situace, ve kterých se standardně používá tento způsob napojování:

a) opakované výzkumy založené na nezávislých výběrech; každých deset let se provádí sociologický výzkum hodnot se stejným dotazníkem, každý měsíc se zjišťují volební preference, důvěra v instituce a politiky nebo se spojují výzkumná data z různých zemí pro komparační analýzu
b) databáze pacientů je pravidelně aktualizována o nové případy minulého týdne
c) tracking – výzkum trhu nebo politických postojů je rozložen do průběhu celého roku, jednotlivé měsíční soubory jsou samostatně zpracovány a poté jsou připojovány ke kumulativně rozšiřovanému celkovému souboru
d) spojováním souborů o prodejích z jednotlivých prodejen, poboček či obchodních zástupců každý týden vzniká úhrnný soubor informací o prodejích organizace

Tip: Při pravidelném a často opakovaném procesu spojování dat doporučujeme zápis rutinní procedury připojování ve formě syntaxe a automatizované spouštění akce v rámci Windows.

B) Připojení souboru nových proměnných s (částečně) stejnými případy k pracovnímu souboru

`Data – Merge Files – Add Variables`

Připojení datových souborů *„do šířky"*, tj. přidávání proměnných k aktivnímu pracovnímu datasetu, předpokládá navázání případů obou souborů, proto mezi soubory musí být určena korespondence, aby akce byla vůbec smysluplná. Pro zcela shodné soubory, ve kterých jsou případy stejně uspořádány, si řádky korespondují automaticky. Pokud v souborech máme řádky, které nejsou jednoznačně párovány a uspořádány, používáme pro přiřazení *klíčovou proměnnou (key variable),* která spojení jednoznačně určuje. Klíčových proměnných může být více, přiřazení je dáno jejich kombinací.

Spojení je řízeno posloupností kroků podle návodných oken.

Před spojováním souborů zajistíme v obou souborech seřazení případů podle stejného hlediska klíčových proměnných (identifikace).

B1) Spojení dvou souborů se stejnými případy

1. Určení výchozího datasetu, ke kterému bude připojen jiný soubor; je jím aktivní otevřený soubor
2. Určení souboru, který bude připojen v okně ***Add Variables to „název aktivního datasetu"***:
   a) potvrdíme jeden z otevřených datasetů v konkrétní analytické situaci (jsou všechny vyjmenovány v okně) nebo
   b) lokalizujeme a otevřeme externí soubor *.sav* (připojení souborů jiných typů než *.sav* musí předcházet jejich otevření ve formátu *.sav*)
3. V dalším okně ***Add Variables From*** *„název připojovaného datasetu"* jsou soupisy
   a) všech proměnných, které budou obsaženy v odvozeném souboru (***New Active Dataset***) a
   b) proměnných, které jsou ze spojení vynechány, tj. mají stejná jména v obou souborech (***Excluded Variables***); proměnné aktivního souboru jsou základem pro připojení
      i) proměnné z aktivního jsou označeny (*), proměnné v napojovaném jsou značeny (+)
4. Označením nepoužité proměnné a tlačítkem ***Rename*** lze změnit jméno proměnné (a tak zajistit připojení)

5. Zaškrtnutím a určením názvu (***Indicate case source as variable***) zavedeme do odvozeného souboru dodatečnou proměnnou indikující zdroj údaje v řádku (0 = aktivní soubor, 1 = připojovaný soubor)
6. Potvrdíme nebo určíme klíčovou proměnnou
   a) v případě stejně velkých souborů se stejnými a shodně uspořádanými případy stačí potvrdit akci tlačítkem ***OK***
   b) nejsou-li soubory totožné, použijeme klíčové proměnné
7. Zatrhneme volbu ***Match cases on key variables***
8. Zatrhneme volbu ***Cases are sorted …***
9. Zvolíme ***Both files provide cases***
10. Určíme *klíčové proměnné*, tj. proměnné, které jednoznačně určují případyproměnné převedeme do okna ***Key Variables***, klíčové proměnné se musejí vyskytovat v obou souborech pod stejným jménem. K sjednocení jmen případně použijeme tlačítko ***Rename***.
11. Potvrdíme ***OK***

Příklad 1.8: *Komparace výsledků voleb v roce 2010 a 2013 v okresech*

Spojovat budeme soubory *„okresy 2010 - volby.sav“* a *„okresy 2013 - volby.sav“*. Zvolíme jeden ze souborů, např. rok 2013, jako aktivní a druhý připojíme. K tomu ale musí být srovnatelné proměnné nazvány různě – např. *„čssd_10“* a *„čssd_13“*, *„ods_10“*, *„ods_13“*, …proto před vlastním spojováním zkontrolujeme jména a případně provedeme přejmenování. Klíč ke spojení je proměnná *„okres“*. Proměnné ze souboru roku 2010 jsou připojeny na konci souboru 2013. Tím můžeme párově porovnávat změny u jednotlivých stran např. párovým t-testem.

Příklad 1.9: *Klasifikace studentů a výsledky dotazníků o zájmech*

Ve třech souborech jsou pro studenty jedné třídy uloženy výsledky ústních zkoušek z matematiky, ve druhém souboru jsou uloženy výsledky testů a třetí soubor obsahuje záznamy odpovědí na dotazník o zájmech a o představách o budoucnosti studentů. Ve všech souborech jsou případy identifikovány proměnnou ID, která slouží jako klíčová pro případ, že soubory nejsou úplné. Sloučení souborů se provádí postupně: Nejprve určíme jeden ze souborů jako aktivní dataset (ústní zkouška), k němu připojíme jeden z dalších (výsledky testu) a k výsledku připojíme třetí soubor (dotazník).

B2) Připojení informace k danému aktivnímu souboru podle číselníku (klíčových proměnných)

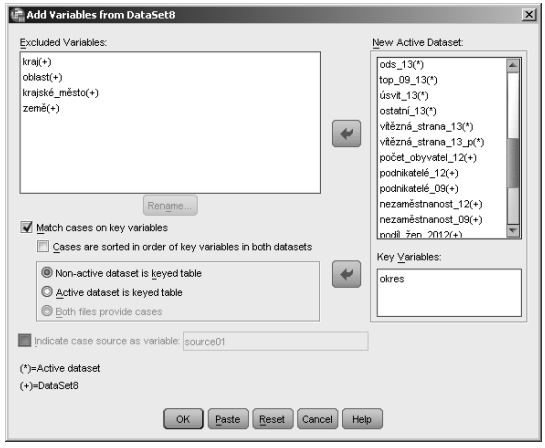
Kromě spojování souborů se vzájemně odpovídajícími si prvky můžeme stejným způsobem přidávat informaci z jiných zdrojů podle zvolených číselníků.

Zajistíme setřídění obou souborů podle klíčové proměnné, kterou se spojení naváže. Aktivizujeme základní soubor, ke kterému budeme připojovat informaci, a provedeme prvních osm kroků stejně jako v případě B1. Pak pokračujeme:

9. Zvolíme ***Non-active dataset is keyed variable***.
10. Určíme klíčovou proměnnou, tj. proměnnou, podle které bude napojení provedeno. Je společná oběma souborům a má v obou souborech stejné jméno, jednoznačně přiřadí informaci k aktivnímu datasetu. Určíme ji převodem do okna ***Key Variables***.

11. Potvrdíme ***OK***.

Klíčových proměnných v kroku 10 může být více a mohou na sebe v kombinaci svých hodnot navazovat.



Obrázek 1.16 Zadávací okno pro připojení proměnných – napojení souborů se stejnými případy, ale s novými proměnnými podle klíče okres – k okresním volebním výsledkům se připojují sociálně-demografické charakteristiky okresu

Příklad 1.10: *Spojená informace o pacientech*

K jednotlivým záznamům o opakovaných návštěvách u lékaře (řádky v souboru) připojíme informace o věku, prodělaných chorobách, chorobách rodičů apod. každého z pacientů (řádky v jiném souboru), které jsou na jedné konkrétní návštěvě nezávislé a byly získány jako soubor vstupních záznamů.

Příklad 1.11: *Komparace okresů a kraje u volebních výsledků*

Chceme-li komparovat výsledky okresů s výsledky jejich krajů a máme-li okresní a krajské výsledky v samostatných souborech, můžeme krajské výsledky napojit, abychom mohli data zpracovávat v rozdílech mezi okresy a jejich kraji.

Připojovat můžeme i soubor s jedním řádkem, klíčovou proměnnou pak je stejná konstanta v obou souborech.

Agregace případů

`Data – Aggregate`

V mnoha datových analýzách vystupují jednotky různých stupňů, které vznikají postupným hierarchickým spojováním do skupin s nějakou společnou vlastností. Lidé vytvářejí rodiny, rodiny žijí v jedné obci, obce se spojují kolem střediskové obce atd. Studenti středních škol patří do jednotlivých tříd, třídy tvoří školu, školy patří k nějakému územnímu celku. Volební okrsky patří k obci, poté k volebnímu kraji.

Takováto seskupení jsou dána přirozenou cestou běžné praxe, mohou být výsledkem různých procesů, jsou dána výzkumem (klasifikace diagnóz v medicíně, taxony v biologii) nebo pragmaticky dlouhou zkušeností či vznikají za nějakým účelem. Vznikají také analyticky pomocí různých metod analýzy, typicky seskupovací analýzou, nebo určením kombinace třídících, resp. klasifikačních, znaků. V analýze dat takové skupiny zpracováváme komparačně a jako faktory nebo efekty v kauzálních vztazích nejrůznějšími statistickými způsoby (k tomu slouží metody Části 2 této knihy).

Jsou-li známy jednotlivé případy, z nichž skupinky vznikají, musíme je nejprve agregovat. Pro souhrnné statistické zpracování takovýchto jednotek potřebujeme zavést jejich soubor, v němž řádek reprezentuje charakteristiku agregované skupinky. Data o takovýchto agregátech získáváme přímo z vnějších zdrojů, avšak často je můžeme (resp. musíme) získat odvozeně tím, že spojíme případy do skupin statistickou sumarizací z údajů o jejích členech. Výsledek použijeme jako nové záznamy pro další analýzu.

V praxi to znamená, že u jednotlivých proměnných původního souboru odvodíme z individuálních údajů charakteristiku skupinky a tak vytvoříme novou proměnnou pro soubor nových agregovaných jednotek. Agregace znamená výpočet nějaké statistické charakteristiky (nebo statistik) pro každou skupinku. Agregované statistiky jsou buď připojeny ke všem jednotkám skupiny, nebo vytvoří odvozené soubory, které jsou zpracovány v průběhu analýzy zvlášť nebo uloženy pro pozdější samostatnou analýzu.

Jsou to na příklad příjem domácnosti, počet členů rodiny, charakteristika typu rodiny, průměrná známka z matematiky ve třídě, celkový průměrný úspěch za celou školu, počty, procenta, minimální hodnoty, maximální hodnoty, rozpětí zisků politických stran v rámci volebního okrsku, okresu, kraje, regionu apod. Charakteristiky vznikají výpočtem či odvozením z jednotlivých údajů jednotlivců, resp. z údajů skupin nižší úrovně.

V **IBM SPSS Statistics** se vytváří soubory agregovaných jednotek sloučením případů do nových případů (skupinek) a výpočtem zvolených charakteristik v proceduře **Aggregate**. V ní definujeme skupinky pomocí třídicích proměnných a jejich kombinací a sloupce pomocí některé ze statistických funkcí, které jsou k dispozici v proceduře. (Poznamenejme také, že agregované soubory lze v průběhu analýzy tvořit také z výstupů analytických procedur jejich převedením na **IBM SPSS Statistics** soubor přímým uložením, kopírováním do nového prázdného datasetu nebo procedurou **OMS**.)

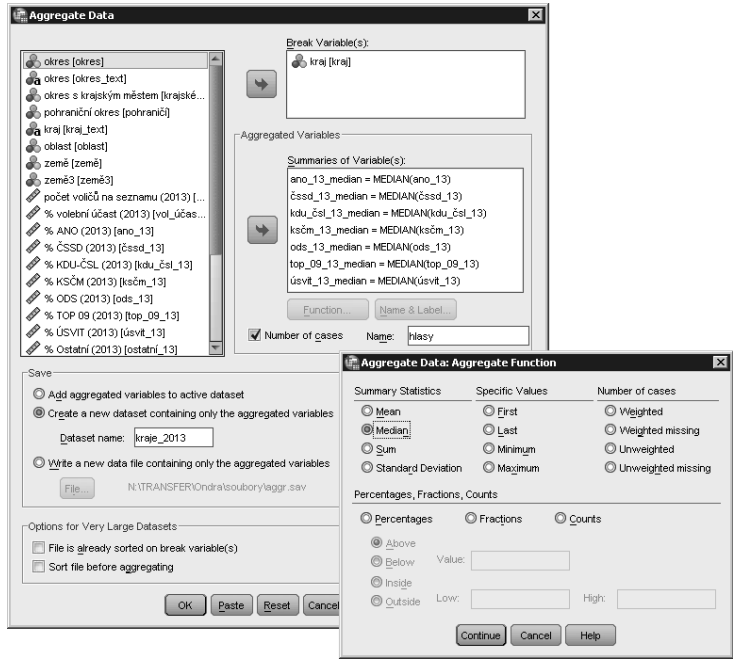
Pro postup agregace volíme **Data – Aggregate**. V okně této volby určíme všechny atributy agregačního postupu a výsledného souboru agregovaných jednotek.

Postup zadání:

1. Proměnnou, která definuje skupinky, převedeme do okna ***Break Variable(s)***; tato proměnná svými hodnotami určuje agregační skupiny. Do okna lze převést více proměnných, skupiny jsou pak určeny kombinací jejich hodnot.
2. Nové proměnné jsou určeny v rámečku ***Aggregated Variables***. Agregované hodnoty budou zapsány v nových sloupcích a budou vycházet z proměnných převedených do okna ***Summaries of Variable(s)***. Z jedné proměnné původních případů lze vytvořit více proměnných tím, že ji převedeme vícekrát a pak předvolenou agregační funkci změníme.
3. Po převedení proměnných se jejich jména změní pro agregovaný soubor podle předvolby: ke jménům původních proměnných se přidá text „*_mean_1 = MEAN(„původní jméno“)*“.

Část před rovnítkem je jméno agregované proměnné, v němž *mean* značí způsob jejího vzniku (výpočet charakteristiky) a číslovka *1* značí, že jde o první použití proměnné pro nový soubor. Předvolbou je nabídka průměru hodnot. Po označení jedné, několika vybraných či všech proměnných (při převádění jsou označeny automaticky všechny) můžeme u nich měnit tuto funkci na jinou ze seznamu. To provedeme v tlačítku ***Function***, které se zároveň s převodem aktivuje a v němž lze změnit průměr na jinou agregační funkci. Tyto funkce jsou uvedeny v okně, které je vyobrazeno na obrázku 1.17.

4. Při zatržení pouze jedné proměnné v okně se aktivuje tlačítko ***Name&Label***, v němž můžeme změnit její jméno a zavést pro ni název.
5. Zatržením volby ***Number of cases*** vyžádáme novou proměnnou, která bude značit počet původních případů ve skupině a i změnit její předvolené jméno (*N_BREAK*).
6. V rámečku ***Save*** určíme formu výstupu agregovaných hodnot. K dispozici jsou tři alternativy:
   - ***Připojení v aktivním souboru – Add aggregated variables in active dataset*** přiřadí agregované hodnoty všem případům agregované skupiny v aktivním, výchozím, datasetu.
   - ***Nově otevřený dataset – Create a new dataset containing only aggregated variables*** vytvoří nový dataset, jehož řádky budou odpovídat agregovaným skupinám a jehož sloupce budou reprezentovat pouze proměnné vytvořené v rámečku ***Save***. K tomu musíme zadat jméno nového datasetu v ***Dataset name***.
   - ***Nový uložený soubor – Write a new data file containing only aggregated variables*** uloží vzniklý soubor v *.sav* formátu na určenou adresu v tlačítku ***File…***
7. Pro velké soubory jsou k dispozici ještě volby pro setřídění souboru.



Obrázek 1.17 Tabulka agregačních funkcí – agregace okresních volebních výsledků na krajskou úroveň, s přepnutím předvolby (průměr) na medián

Při agregaci je důležité také rozhodnout, zda statistiky budou počítány váženě (např. svojí velikostí), nebo přímo, tak jak jsou zaznamenány bez vah. To je dáno procedurou **Weight** zapojenou před agregací. Vážené charakteristiky můžeme vyžadovat, např. budou-li agregovány okresní volební výsledky na krajskou úroveň. Žádaný výsledek dostaneme kombinací procedur **Weight** (viz kap. 2) a **Aggregate.**

Tip: Agregovat můžete i na celý soubor. K tomu stačí nechat pole ***Break Variable(s)*** prázdné. Takový postup je vhodný k tomu, abychom agregované proměnné připojili k souboru a poté zjistili odchylky původních proměnných od ní.

Příklad 1.12: *Agregace okresních volebních výsledků na krajské*

V souboru „*okresy 2013 - volby.sav*" jsou uloženy výsledky volebních zisků pro jednotlivé okresy. Převod na agregovaná data pro kraje provedeme postupně podle bodů uvedených výše. Podle cíle analýzy musíme rozhodnout, zda budeme chtít při agregaci jednotky (okresy) vážit jejich mírou velikosti (v našem případě počtem platných hlasů – proměnná „*hlasy _13*") nebo zda budeme považovat všechny okresy za stejně důležité jednotky a necháme jim předvolenou váhu 1. Rozhodnutí o váze je rozhodnutím analytickým, stejně tak jako volba váhy. Mírou velikosti skupinky používanou jako váha může ale být i počet voličů na seznamu (proměnná *„voliči_v_seznamu_13")* či počet obyvatel (v souboru není k dispozici, bylo by nutné proměnnou přidat manuálně nebo příkazem **Merge – Add Variables**). Postup vážení bude popsán v kapitole 2. V našem případě je to **Data – Weight Cases –** volba ***Weight cases by*** – převedení proměnné „*hlasy_13*" do okna ***Frequency Variable***.

Rozhodnutí o vahách záleží na konkrétní úloze. U agregace procent průměrem je v mnoha případech vážení velikostí jednotky logické, u agregace součtem naopak dostaneme vážením chybné výsledky. Vážit lze ovšem i designovými vahami pravděpodobnostního výběrového postupu nebo nějakou další proměnnou vyjadřující např. důležitost jednotky. Rozhodnutí o vahách je dáno také povahou dat. V souboru okrskových výsledků voleb za roky 2010 a 2013 můžeme pro společnou analýzu vážit buď počtem voličů z jednoho, nebo z druhého roku. Přesným postupem by bylo rozdělit soubor na dva pro každý rok zvlášť (vzhledem k různým počtům aktivních voličů), agregovat je jednotlivě a poté je spojit procedurou ***Data – Merge Files – Add Variables***. Tak bychom dostali správné údaje pro vyšší agregované celky.

Pro vlastní agregaci v **Data – Aggregate** určíme ***Break Variable*** jako „*kraj*", do ***Summaries of Variables*** převedeme sedm proměnných „*ano_13*" až „*úsvit_13*", agregované hodnoty budou průměry. V obr. 1.17 jsou označeny všechny agregované proměnné a v tlačítku ***Function*** je přepnut předvolený průměr (***Mean***) na ***Median***. Vyžádáme proměnnou vyjadřující počet hlasů v kraji v ***Number of cases***, kterou nazveme „*hlasy*" (počet případů se počítá jako součet vah). V ***Save*** zvolíme výstup v novém datasetu, který pojmenujeme „*kraje_2013*". Potvrdíme ***OK***. U každé strany můžeme agregací získat několik charakteristik: průměry, mediány, maximální a minimální hodnoty okresních volebních výsledků v rámci kraje apod.

KAPITOLA 2

Případy

V této kapitole:

- Manuální úpravy
- Uspořádání případů
- Výběr případů – práce s podmnožinou záznamů
- Štěpení souboru
- Vážení

Manipulace s případy, tj. s řádky datové matice, zahrnují základní aktivity, které směřují

a) k doplnění či opravě dat
b) k čištění dat a k první analytické informaci odhalující nekvalitu a nutnost zásahu
c) k přípravě vhodné formy dat pro analýzu

K tomu slouží:

- uspořádání případů
- výpisy dat a základní sumární popis
- výběry podmnožin pro analýzy na podsouborech
- štěpení souboru na části pro paralelní analýzu
- vážení případů
- odstranění duplikátů

Procedury pro práci s řádky nalezneme v hlavním menu **Data** nebo je provádíme přímo v datové matici.

Manuální úpravy

V okně **Data View** lze přímo a ručně provádět některé úpravy. Především je to prostá změna zápisu hodnoty v poli. Tu provádíme obvykle v případech, kdy opravujeme chybné hodnoty či doplňujeme chybějící údaje. Akci provádíme obdobně jako např. v MS Excelu: označíme si určené pole a v něm hodnotu přepíšeme nebo zapíšeme.

Do jednotlivých polí, řádků, sloupců nebo bloků polí můžeme také kopírovat hodnoty odjinud (z vnějšího zdroje nebo z jiného pole). Tím se původní hodnoty přepíší. Celé řádky se odstraní buď označením a tlačítkem ***Delete***, nebo označením, kliknutím pravým tlačítkem myši a volbou ***Clear***.

Vkládání nových prázdných případů nabízí dvě cesty. Označíme řádek, na němž má vzniknout nový případ, a buď

a) potvrdíme v menu **Edit** volbu **Insert Cases**, nebo
b) otevřeme menu pro řádek klepnutím pravým tlačítkem myši na čísle řádku a potvrdíme stejnou volbu **Insert Cases**.

Tento postup se používá ke kopírování řádků: *Vytvoříme prázdný řádek a pak do něj zkopírujeme přenos*.

Případy lze také vyhledávat (**`Edit – Find`**) a nahrazovat (**`Edit – Replace`**). K rychlému přeskoku na konkrétní řádek použijeme **`Edit – Go to –`** ***Case***, zapíšeme *číslo řádku* a potvrdíme ***Go***.

Změnu pozice řádku lze provést přetažením myší: Označíme číslo řádku v prvním levém sloupci a se stisknutým levým tlačítkem myši jej a přesuneme na žádané místo (podle indikace červené čáry, ukazující polohu přenášení). Pro rychlý přenos na začátek stačí přejet s případem nad matici dat na lištu okna. (Zrychlený postup je vhodný především pro velké soubory s velkým počtem případů.) Tímto způsobem přenášíme současně i více označených případů (několik případů za sebou označíme prostým tahem myši nebo potvrzováním případů se stisknutou klávesou ***Ctrl***).

Uspořádání případů

`Data – Sort Cases`

Uspořádání případů je důležitou funkcí ze dvou důvodů:

a) Je to rychlá kontrola nejmenších a největších hodnot a odhalení extrémních hodnot k posouzení jejich validity, správného zápisu a meritorní příslušnosti k souboru.
b) Je to první informace o datech identifikací minima, maxima, a tím i rozpětí dat; jde tedy o první (a nejjednodušší) analytický krok.

Uspořádat data ve sloupci můžeme vzestupně nebo sestupně. Můžeme také uspořádat případy postupně podle více kritérií. K dispozici jsou dva přístupy:

- *Rychlý přímý (zjednodušený) postup v datové matici:* označíme sloupec v **Data View**, otevřeme menu kliknutím na záhlaví sloupce a potvrdíme buď vzestupné (***Sort Ascending***), nebo sestupné (***Sort Descending***) přeuspořádání případů. Chceme-li uspořádat případy podle více hledisek, označíme (pomocí ***Ctrl***) příslušné sloupce a ty se pak seřadí tak, že nejprve se provede řazení prvního sloupce odleva a poté postupně uvnitř jednotlivých stejných hodnot tohoto sloupce se řadí případy podle druhého a tak postupně dále. Všechny sloupce se řadí stejným směrem; pro jiné pořadí sloupců pro uspořádávání je třeba je přeřadit (tahem myši) v datové matici.
- *Procedura uspořádání v menu* **`Data – Sort Cases`***:* komplexnější uspořádání provedeme převedením řadicích kritérií do ***Sort by*** postupně podle hierarchie uspořádání a postupně u každého určíme, zda půjde o sestupné nebo vzestupné řazení: označíme kritérium (název sloupce) a určíme ***Sort Order***. (Uspořádanou datovou matici můžeme uložit.)

Tip: Přímým postupem v datovém okně můžete provést i složitější uspořádání dat tím, že postup rozložíte na kroky: nejprve realizujte uspořádání posledního (nejpodrobnějšího) kritéria podle zvoleného směru, pak pokračujte k druhému atd.

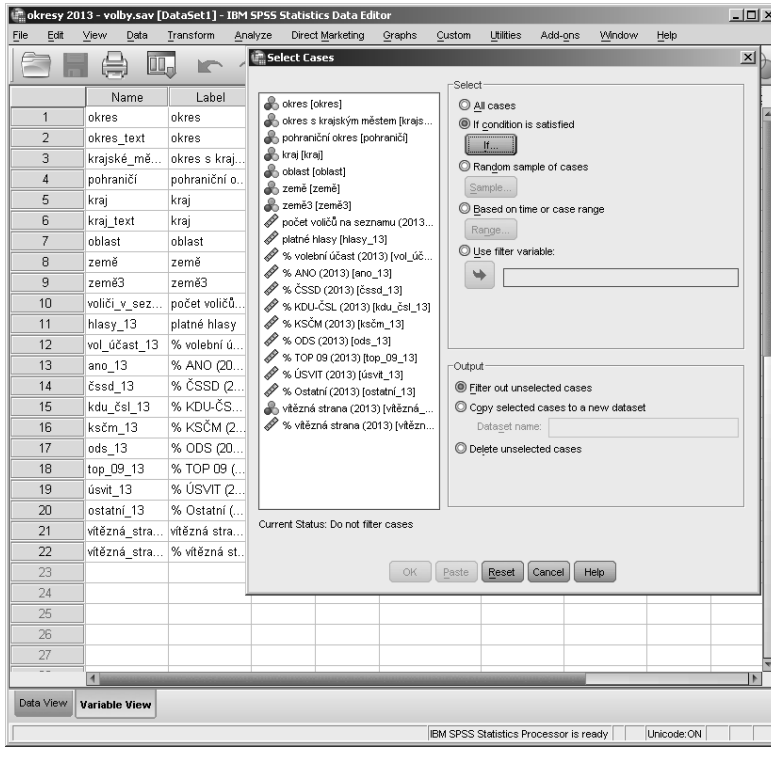
Příklad 2.1: *Kraje a volební účast*

1. V souboru „*okresy 2013 - volby.sav*“ seřadíme okresy podle hodnot volební účasti sestupně (proměnná „*vol_účast_13*“) nejrychleji tak, že v názvu sloupce otevřeme pravým tlačítkem myši nabídku a potvrdíme **Sort Descending** – akce se okamžitě provede;
2. Seřazení podle krajů (vzestupně) a v nich podle volební účasti (sestupně) provedeme jedním ze dvou postupů:
   a) Z menu **`Data – Sort Cases`** tak, že převedeme do okna **Sort by** nejprve proměnnou „*kraj*“ a pak na druhé místo převedeme „*vol_účast_13*“. *Za* oběma proměnnými se napíše „*Ascending*”, což znamená, že předvolbou řadíme případy u obou kritérií vzestupně. Protože chceme ale volební účast v sestupném pořadí, označíme si myší proměnnou „*vol_účast_13*“ *v* okně **Sort by** a přepneme na **Descending** v **Sort Order**. Tím se přepíše *(Ascending)* na *(Descending)*. Potvrdíme a akce se provede.
   b) Přímo v datovém okně získáme tentýž výsledek: nejprve seřadíme případy podle „*vol_účast_13*“ sestupně pomocí nabídky sloupce a v druhém kroku seřadíme obdobně data podle „*kraj*“ vzestupně.
3. Kdybychom chtěli řadit přímo v datové matici případy podle volební účasti v jednotlivých krajích, přičemž by byla obě kritéria byla souladná (vzestupná nebo sestupná), označili bychom oba sloupce (pomocí myši a tlačítka **Ctrl**) a vybrali bychom kritérium v nabídce záhlaví jednoho ze sloupců.

Výběr případů – práce s podmnožinou záznamů

`Data – Select Cases`

Analýza částí datových souborů je v praxi běžná, a proto program poskytuje řadu postupů, jak vybrat žádanou podmnožinu. Výběry realizujeme v proceduře **`Data – Select Cases`**.



Obrázek 2.1 Nabídka výběru podmnožiny – způsob vybírání a tvar výstupu

Nabídkové okno poskytuje několik způsobů vzniku podmnožiny záznamů podle požadavků analýzy v části ***Select***:

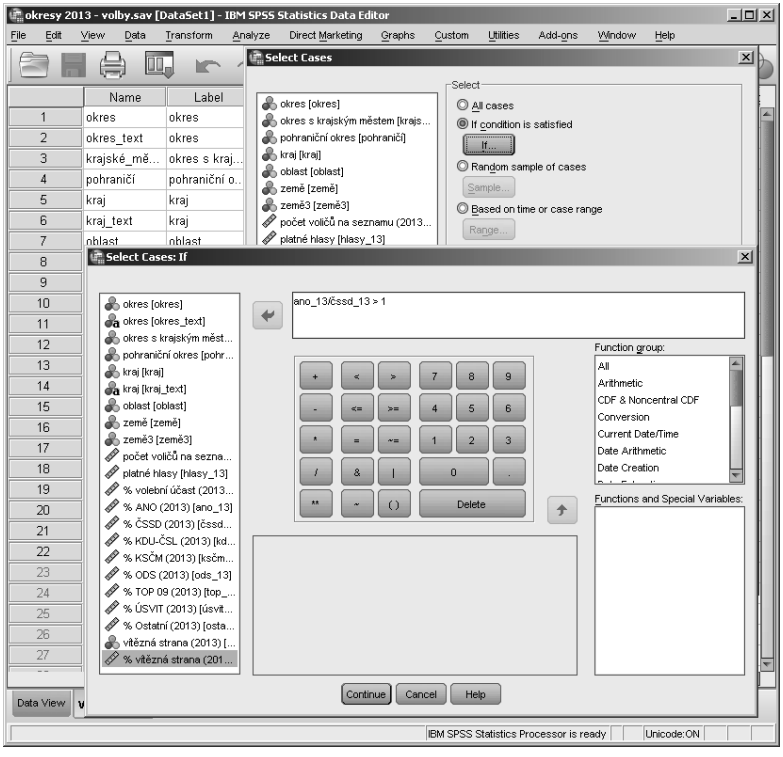
a) *Výběr se neprovádí* – **All cases**, používáme jako předvolbu a pro zrušení výběru, který již byl zaveden.

b) Výběr podmínkou, která vychází z informace v souboru a z náhodných čísel: **If condition is satisfied**. Volba otevře tlačítko **If**, které obsahuje kalkulačku pro určení podmínky výběru.

c) *Pravděpodobnostní výběr*: **Random sample of cases**. V nabídce tlačítka **Sample** zvolíme alternativu přibližného nebo přesného výběrového rozsahu a doplníme požadavek – přibližné procento případů (**Approximately __ % of all cases**) nebo přesný počet případů ze zvoleného počtu prvních případů (**Exactly__cases from the first __ cases**); přirozenou volbou druhého parametru je velikost souboru.

d) Výběr úseku časové (nebo jinak uspořádané) řady případů**: Based on time or case range**. Tlačítko **Range** otevře okno pro určení čísla prvního a posledního případu (**Observation: First Case __ Last Case __**).

e) Výběr podle určené filtrační proměnné typu (0 = vynech, 1 = vyber): **Use filter variable**. Do okénka volby se převede proměnná souboru, která bude filtrovat případy.



Obrázek 2.2 Podmínka pro výběr vychází z proměnných souboru, z funkcí dat, z aritmetických a logických operací a z náhodných čísel generovaných kalkulačkou

Výběr podmínkou je určen kalkulačkou. Možnosti vytváření podmínky jsou velmi bohaté a využívají stejné funkce a operace jako při tvorbě nových proměnných (viz kap. 3). Všechny funkce a operace jsou podrobněji uvedeny v Apendixu B.

Při zadávání výběru podmnožiny ještě určíme v okně ***Output*** jeden ze tří tvarů výstupu:

a) **Filter out unselected cases.** Případy jsou filtrovány, ale zůstávají v datové matici beze změny. Vynechané řádky poznáme tak, že číslo řádku je přeškrtnuto. Zároveň se vytvoří filtrační proměnná *„filter_$“*, která se stává součástí souboru. I po zrušení výběru přechodem na **All cases** tato proměnná zůstane zachována, takže se k výběru můžeme kdykoliv později vrátit. Při dalším výběru je tato proměnná přepsána. V obou základních oknech programu se vpravo dole na liště objeví zápis *Filter On*, který zmizí při zrušení výběru.

Tip: Provádíte-li výběry vícekrát, je vhodné proměnnou *„filter_$“* vždy přejmenovat, aby zůstala k dispozici i v dalších krocích práce. Přejmenování je užitečné také proto, že charakterizuje význam a genezi podmnožiny, a ulehčuje tak další využití. Pokud takto vzniklý filtr dále nepotřebujete, proměnnou jednoduše zrušte.

b) **Copy selected cases to a new dataset.** Určíme název datasetu v **Dataset name**. Příkaz ponechá původní soubor tak, jak byl, a vybrané případy, tj. redukovaný soubor, umístí do nového datasetu.

c) **Delete unselected cases.** Aktivní pracovní soubor je redukován, nevybrané případy se odstraní. Při této volbě lze doporučit opatrnost, protože ztrácíme informaci, která v daném souboru již později není k dispozici.

Tip: Před prací s podsoubory doporučujeme vždy uložit záložní kopii původního souboru.

Tip: V případě, že chceme trvale odfiltrovat jen několik málo případů, je nejjednodušším způsobem odstranění případů manuálně, a to buď v původním datovém okně nebo ve zkopírovaném datasetu (**Copy Dataset**).

Příklad 2.2: *Výběry podsouborů okresů pro dílčí analýzy*

a) V souboru „*Okresy 2013 - volby.sav*“ vyloučíme pražské obvody filtrační podmínkou v kalkulačce

1. Pražské obvody mají v proměnné „*kraj*“ kód 1, a proto podmínka pro vyloučení je určena zápisem kraj ~= 1 (~= znamená „nerovná se“).

2. Podmínku buď
   a) celou zapíšeme manuálně z klávesnice, nebo
   b) proměnnou „*kraj*“ označíme v seznamu a převedeme šipkou nebo tahem myši, znak ~= a číslo 1 potvrdíme na klávesnici nabídkového okna.

3. Protože nechceme ztratit vynechané případy ani zakládat nový dataset, necháme v okně **Output** předvolbu **Filter out unselected cases**.

4. Potvrzením **Continue** a **OK** se provede filtrace, označená jednak zápisem na dolní liště a jednak přeškrtnutím čísel případů odpovídajících pražským obvodům.

b) Ve stejném souboru vybereme všechny okresy, ve kterých bylo ANO úspěšnější než ČSSD:

1. Podmínku vyjádříme tak, že podíl úspěchu ANO a úspěchu ČSSD je větší než 1:
ano_13 / cssd_13 > 1

2. Zapíšeme ji opět buď manuálně, nebo převodem proměnných ze seznamu a znaku nerovnosti a číslovky 1 z nabídkové klávesnice.

c) Ve stejném souboru chceme odfiltrovat všechny pražské obvody („*kraj*“, hodnota 1 = 1) a všechny okresy s krajským městem („*krajské_město*“, hodnota = 1), ve kterých je více než 85 000 voličů platnými hlasy („*hlasy_13*“).

1. Podmínku zapíšeme jako
kraj > 1 & ~ (krajské_město = 1 & hlasy_13 >= 85000)

2. Vybíráme všechny okresy, kromě obvodů Prahy, pro které ale současně neplatí, že v nich jsou krajská města s více než 85 000 hlasy.