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.

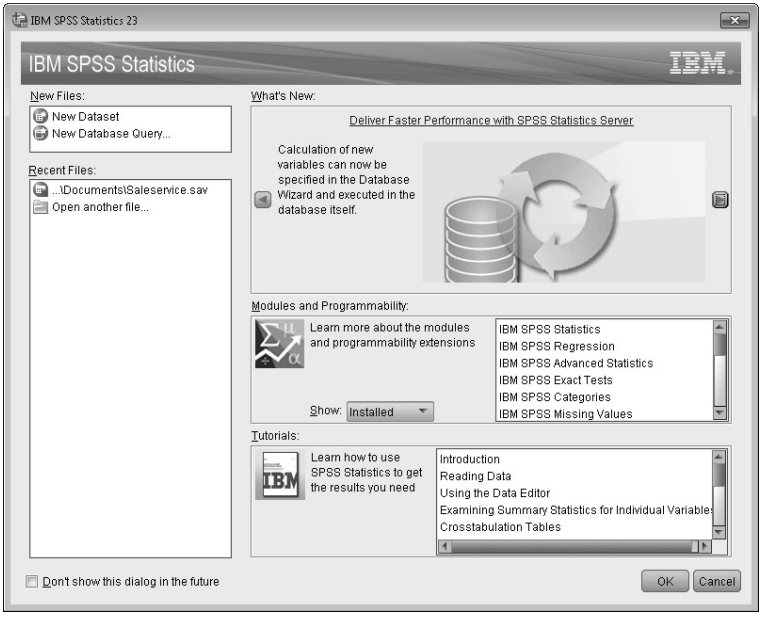
Manuální zápis dat do souboru

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

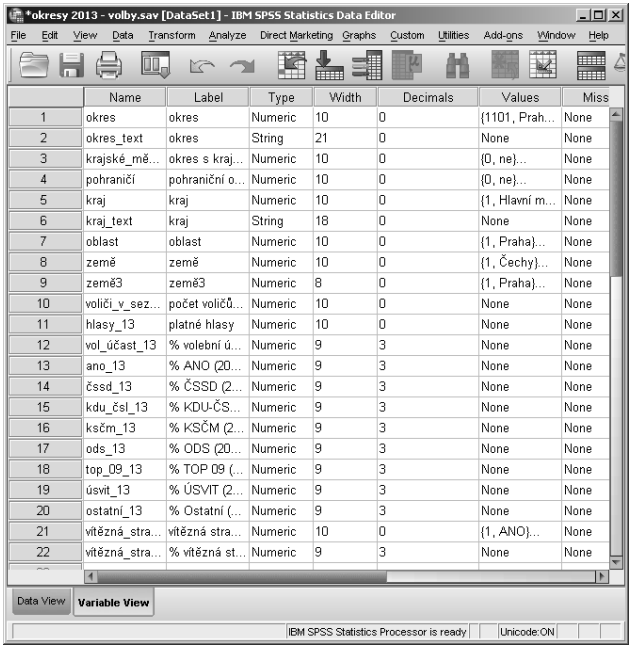
Postup A – při vyvolání programu se otevře vstupní nabídka:

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.

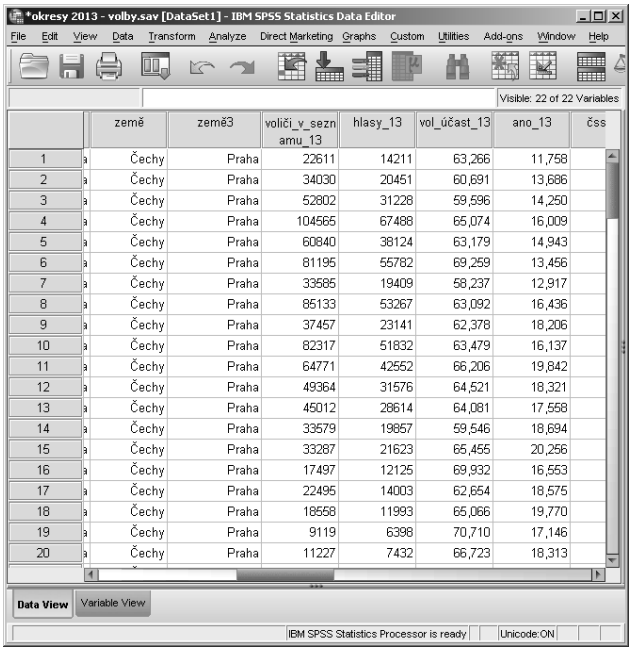


Obrázek 1.1 Okno vstupní nabídky při otevření programu



	Name	Label	Type	Width	Decimals	Values	Miss
1	okres	okres	Numeric	10	0	{1101, Prah...	None
2	okres_text	okres	String	21	0	None	None
3	krajské_mě...	okres s kraj...	Numeric	10	0	{0, ne}...	None
4	pohraničí	pohraniční o...	Numeric	10	0	{0, ne}...	None
5	kraj	kraj	Numeric	10	0	{1, Hlavní m...	None
6	kraj_text	kraj	String	18	0	None	None
7	oblast	oblast	Numeric	10	0	{1, Praha}...	None
8	země	země	Numeric	10	0	{1, Čechy}...	None
9	země3	země3	Numeric	8	0	{1, Praha}...	None
10	voliči_v_sez...	počet voličů...	Numeric	10	0	None	None
11	hlasy_13	platné hlasy	Numeric	10	0	None	None
12	vol_účast_13	% volební ú...	Numeric	9	3	None	None
13	ano_13	% ANO (20...	Numeric	9	3	None	None
14	čssd_13	% ČSSD (2...	Numeric	9	3	None	None
15	kdu_čsl_13	% KDU-ČS...	Numeric	9	3	None	None
16	kscm_13	% KSČM (2...	Numeric	9	3	None	None
17	ods_13	% ODS (20...	Numeric	9	3	None	None
18	top_09_13	% TOP 09 (...	Numeric	9	3	None	None
19	úsvit_13	% ÚSVIT (2...	Numeric	9	3	None	None
20	ostatní_13	% Ostatní (...	Numeric	9	3	None	None
21	vítězná_stra...	vítězná stra...	Numeric	10	0	{1, ANO}...	None
22	vítězná_stra...	% vítězná st...	Numeric	9	3	None	None

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



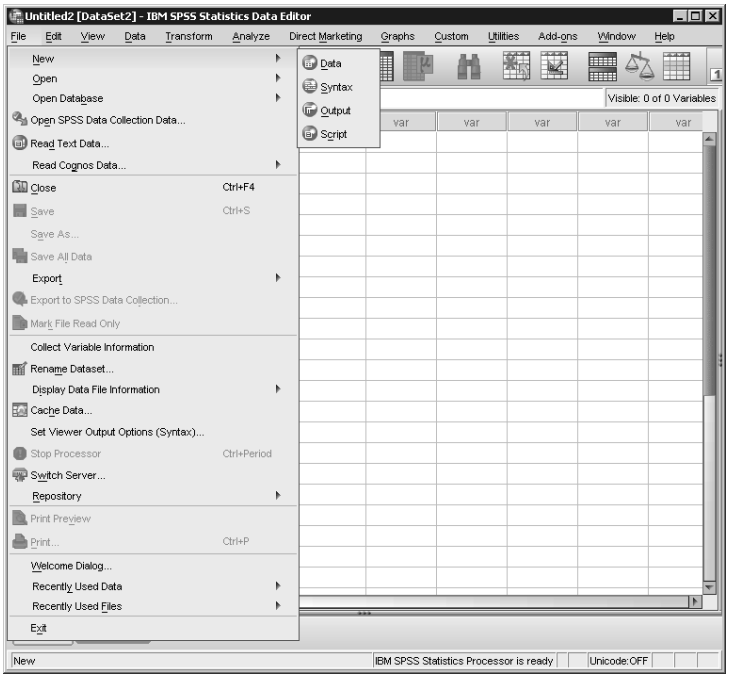
Obrázek 1.3 Datové okno s pořízenými hodnotami

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.

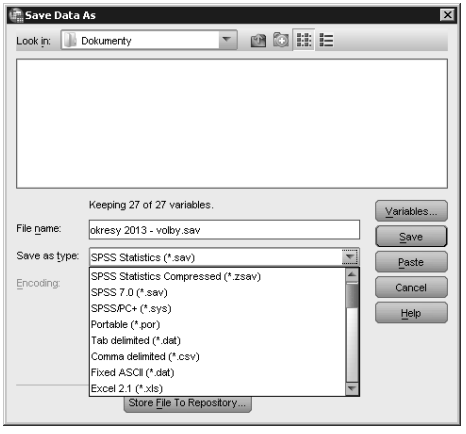
Z hlediska metodiky statistické práce zdůrazňujeme, že všechny nově pořizované soubory musí být – pro zajištění kvality dat i výsledků – nutně zkontrolovány v plném rozsahu všech případů a proměnných.

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:

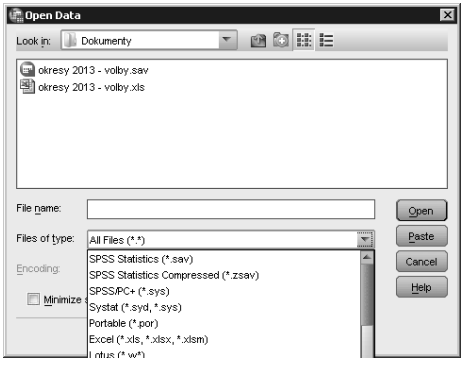
   - soubory Excelu (*.xls*, *.xlsx*, *.xlsm*)
   - textové soubory *(.txt*, *.dat*, *.csv*, *.tab*)
   - soubory dBase (*.dbf*)
   - soubory jiných statistických programů – Stata (*.dta*), SAS (*.sas7bdat*, *.sd7*, *.sd2*, *.ssd01*, *.ssd04*, *.xpt*), Systat (*.sys*, *.syd*),
   - Sylk (*.slk*)
   - Lotus (*.*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.

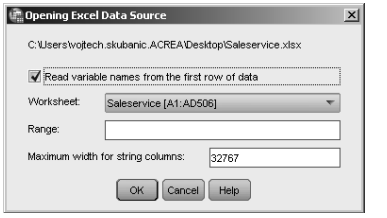
Program s prázdným datovým oknem otevřeme ikonou nebo také potvrzením ze seznamů v obslužných programech Windows či přímo vyvoláním *stats.exe* ze složky *IBM/SPSS/Statistics/23* (resp. číslo instalované verze).

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

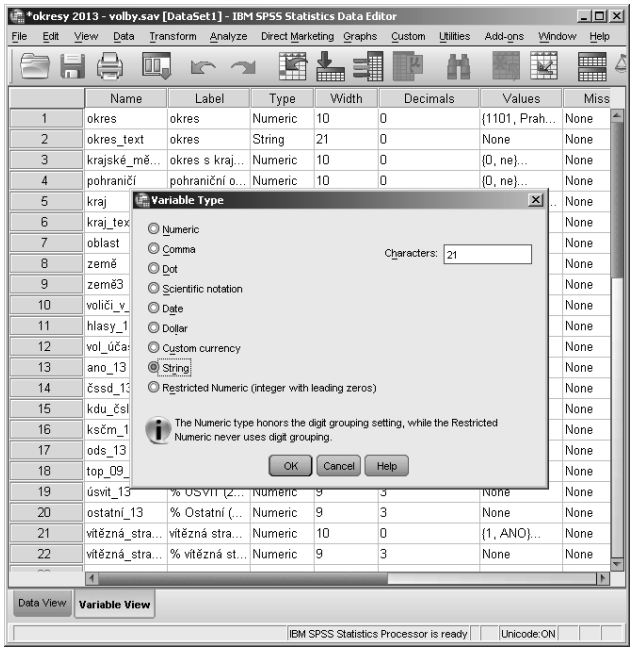
Ze souboru MS Excel tedy přenášíme jen název proměnné a typ proměnné. Musíme dát ale pozor na správné určení prvního řádku – neurčíme-li jej jako řádek s názvy a on přitom názvy obsahuje, program převezme řádek jako datový a určí všechny proměnné jako textové (***String***). Pracovní soubor bude mít 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 …