Kapitola 2

Práce s hromadnými daty před analýzou

2.1 Stručné seznámení s programem IBM SPSS Statistics

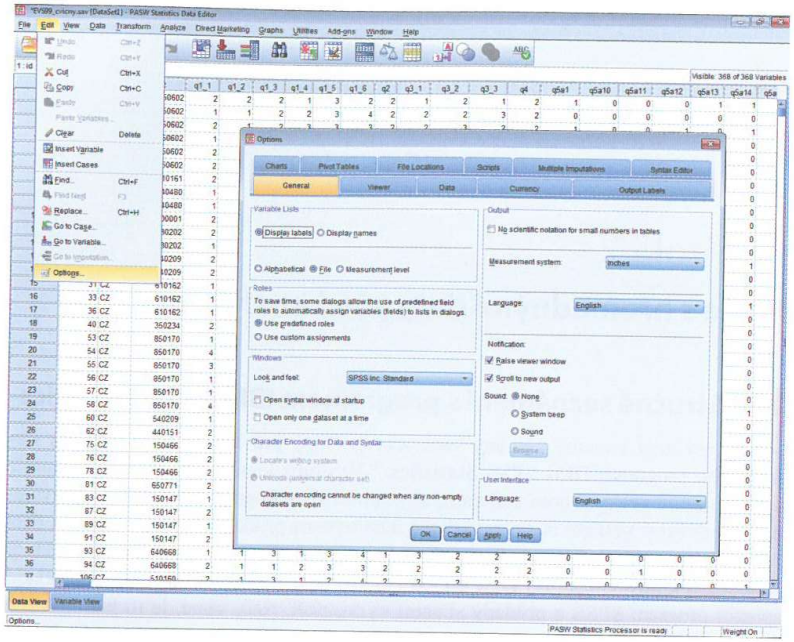
V naší učebnici budeme všechny statistické postupy provádět prostřednictvím statistického programu **IBM SPSS Statistics**.[38] Ve druhé kapitole se proto seznámíme se základními prvky tohoto softwaru, abychom byli schopni jej efektivně a smysluplně využívat.[39] Veškeré postupy budeme ilustrovat na příkladech, které obsahují data z reálných výzkumů. Příslušné datové soubory jsou k dispozici na disku (CD), který je součástí knihy. Čtenářům vřele doporučujeme, aby si při pročítání těchto příkladů spustili program SPSS a příklady si sami vyzkoušeli. Ano, víme, je to pak náročné čtení, ale věřte, skutečně se to vyplatí. SPSS je pouze jeden z řady programů pro statistické zpracování hromadných dat – o jiných statistických balících informujeme v dodatku III. Před zahájením prací v SPSS je vhodné si nejdříve nastavit prostředí programu. Pod tlačítkem *Edit* se skrývá volba *Options* (viz obr. 2.1), kde lze navolit zejména podobu výstupů (výsledků výpočtů) – například grafickou podobu tabulek, popisky proměnných, lze také naučit program správné češtině apod.[40]

---

[38] Tento software má poměrně dlouhou historii. Byl vyvinut ještě v době před existencí osobních počítačů (PC) – v éře velkých sálových počítačů. Jeho původní název zněl prostě: **S**tatistical **P**ackage for **S**ocial **S**ciences neboli SPSS. My se této historie budeme držet a program budeme i nadále pro zjednodušení zkráceně nazývat SPSS.

[39] Pro finální práce na této učebnici (mimochodem vznikala několik let) jsme používali většinou verzi SPSS 18.0, některé úpravy jsme pak zpracovávali ve verzi SPSS 22.0. Zmiňujeme se o tom z toho důvodu, že některé obrázky a výstupy se mohou, pokud bude čtenář používat verze jiné, graficky lehce odlišovat. Ničemu to nevadí, neboť principy a postupy analýzy zůstávají v procedurách, které obsahuje tato učebnice, nezměněny.

[40] Všechny obrázky a výstupy v této kapitole pracují s datovým souborem „EVS99-cvicny.sav".

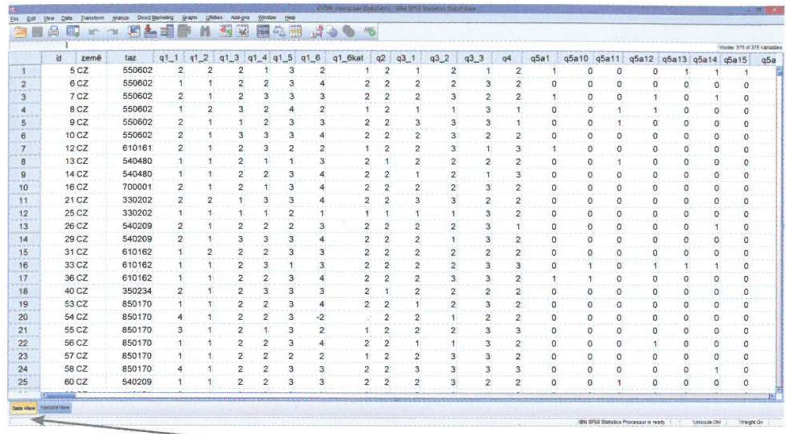


Obr. 2.1 Okno pro nastavení vnitřního prostředí programu

Každá analýza dat začíná nahráváním dat. Abychom data mohli nahrávat, musíme 1) nejdříve definovat jednotlivé proměnné, jimž pak 2) přiřazujeme výzkumem zjištěné konkrétní hodnoty. K operacím 1) a 2) slouží okna *Data View* a *Variable View*. Obě okna jsou interaktivní, takže do nich můžeme psát.

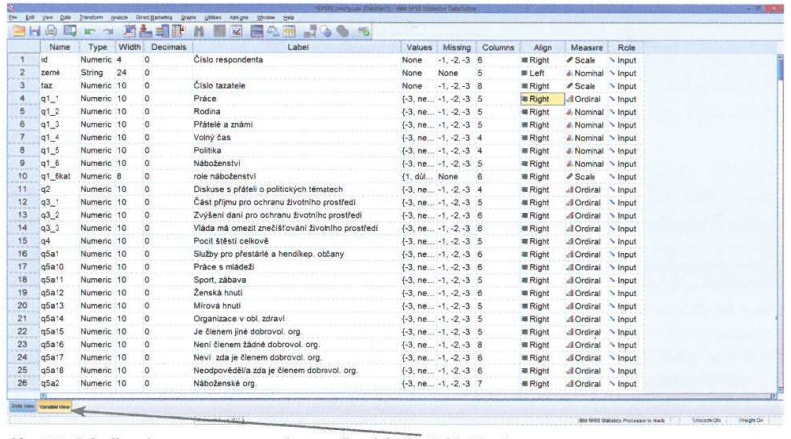
Data View (viz obr. 2.2) obsahuje matici dat, v níž řádky znamenají **případy** (*cases*) výzkumné jednotky – většinou jde o osoby (respondenty), ale výzkumnými jednotkami mohou být i skupiny osob, územní celky, předměty jako texty apod. Sloupce matice jsou **proměnné** neboli charakteristiky těchto zkoumaných jednotek, jejich vlastnosti. Každá jednotka tedy představuje vektor a číslice v něm představují kódy hodnot proměnných (u nominálních a ordinálních proměnných) nebo čísla (u spojité, to je kardinální proměnné) popisující vlastnosti/charakteristiky jednotky.[41]

[41] Pozici řádků a sloupců lze měnit pomocí menu *Data – Transponse*. Děláme to například tehdy, když je výstupní tabulka příliš široká a nevešla by se na šířku tisku. Operace transponování může být užitečná i pro některé pokročilé statistické procedury, např. pro shlukovou analýzu, jak uvidíme v kapitole 14.



Obr. 2.2 Datová matice *(Data View)*

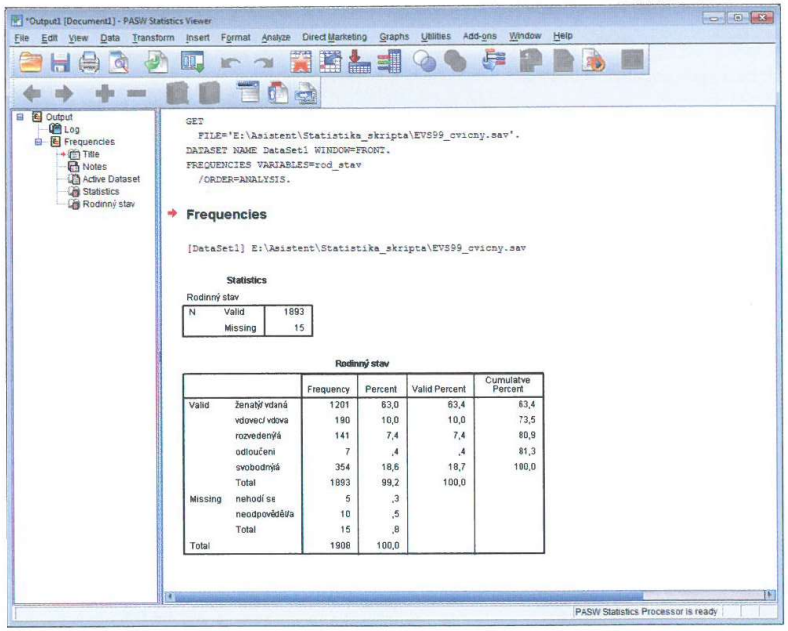
Okno *Variable View* (viz obr 2.3) představuje popis proměnných. Je to v SPSS zabudovaný speciální tabulkový procesor, který tento popis umožňuje. Bez popisu proměnných bychom konkrétní hodnoty proměnných nemohli nahrávat, proto popis proměnných musí vždy předcházet nahrávání dat. Při popisu proměnných vlastně převádíme náš dotazník, jeho jednotlivé otázky či položky, do formalizované podoby, kterou vyžaduje SPSS.



Obr. 2.3 Tabulkový procesor pro popis proměnných *(Variable View)*

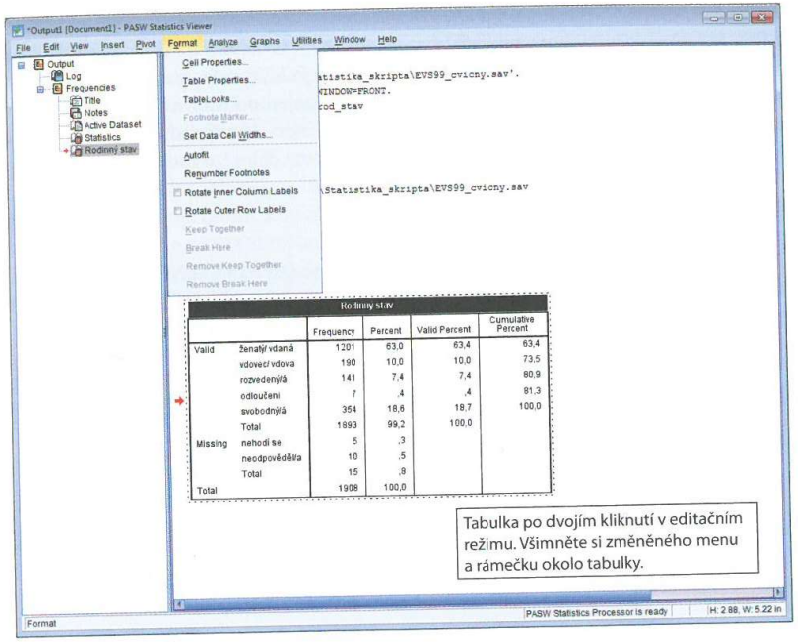
Jednotlivé proměnné, což jsou zkratkovitě vyjádřené jednotlivé otázky z dotazníku, jsou zde – na rozdíl od datové matice – umístěny v řádcích. Sloupce tohoto procesoru pak udávají jejich základní charakteristiky: technické jméno proměnné, její popisek (název), popisky jednotlivých hodnot proměnné, chybějící hodnoty atd.

Třetím základním oknem je okno výstupů, *Output* (viz obr. 2.4a a 2.4b), které se automaticky otevře v okamžiku, kdy zadáme nějaký výpočet. Objevují se v něm výsledky požadovaných výpočtů (tabulky, grafy atd.). Ty zde můžeme editovat.[42] Klikneme-li dvakrát na výstup (tabulku či graf), který chceme editovat, objeví se poněkud jiná nabídka a my můžeme měnit jeho grafickou podobu, měnit texty popisků apod. Postup je naznačen níže v následujících obrázcích. Editovat můžeme především prostřednictvím menu *Edit*, *Format* nebo také *Pivot* (méně často), kde se nabízí zejména již zmíněná a užitečná operace záměny sloupců a řádků.



Obr. 2.4a Ukázka výstupu výpočtu distribuce četnosti (příkaz *Frequencies*)

[42] Pozor, prosím. Pokud máte potřebu editovat své výsledky, veškerou editaci výstupů provádějte zde. Po přenesení výsledků do textového procesoru Word (viz dále) je to již prakticky vyloučeno, a to i v případě, kdy – navzdory varování, která najdete v textu dále – použijete pro přenesení výstupu do Wordu příkazu export. Takto přenesené tabulky/grafy se sice editovat dají, ale při návratu z editačního režimu se tabulka obvykle rozpadne a je nepřehledná až nečitelná.



Obr. 2.4b Editace výstupu

Etická vsuvka: hovoříme-li o editování, máme na mysli pochopitelně pouze editování grafické podoby výstupů. V žádném případě není možné v tabulkách editovat, to je měnit, jejich číselné hodnoty! Obsah výstupů neboli výpočtů z analýz je ve vědě nedotknutelný! Přepis hodnot ve vypočtených tabulkách nebo údajích je ve vědě horším zločinem než plagiarismus. Je to hanebný čin, který má pro jeho aktéra závažné důsledky.

Výsledky, které se objeví v okně *Output*, lze uložit příkazem *Save as*. Uloží se v novém souboru s příponou .spo či .spv.[43] Jednotlivé výstupy i celek lze také exportovat do Wordu, to však nedoporučujeme, neboť tabulky se často rozpadnou. Lepší je v menu *Edit* tabulku zablokovat a pomocí *Copy Object* (nebo také příkazem Ctrl+C) ji vložit jako objekt příkazem Ctrl+V do textu, který píšeme v textovém editoru.

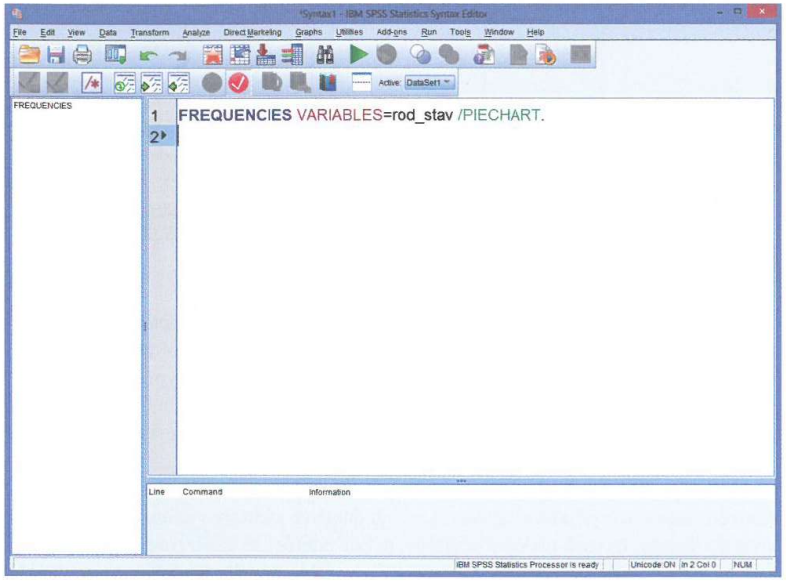
[43] Soubor typu *.spo či *.spv lze otevřít jen v SPSS Statistics. Je nutné také upozornit, že jednotlivé verze SPSS mají různý formát tohoto souboru, a proto lze soubor typu *.spo či *.spv spolehlivě otevřít jen ve verzi SPSS, v níž byl vytvořen. Pro datové soubory (*.sav, viz dále též část 2.4) tato nepřenositelnost mezi verzemi neplatí, nejlepší je ovšem data ukládat ve formátu *.por, který umí spolehlivě číst všechny verze SPSS a nadto i mnohé jiné softwary (např. SAS, STATA Transfer apod.).

Syntax a Script

SPSS byl původně vyvinut v operačním systému DOS, takže místo klikáním na tlačítka v menu bylo nutno zadávat příkazy ve formě vět, jejichž tvar (syntax) byl předepsán. I ve verzi pracující pod Windows lze příkaz zadat nejen pomocí menu, ale i pomocí psaného příkazu. K tomu slouží okno *Syntax* (menu *Files*). Například výpočet rozložení hodnot proměnné rod_stav a grafu tohoto rozložení (viz dále) lze zadat příkazem:

```
FREQUENCIES VARIABLES=rod_stav /PIECHART.          (viz obr. 2.5)
```

Tento příkaz znamená: udělej třídění prvního stupně (=FREQUENCIES) proměnné, která se jmenuje „rod_stav" (neboli rodinný stav respondenta), a přidej koláčový graf (=PIECHART) znázorňující rozložení jejích hodnot.[44]



Obr. 2.5 Okno pro psaní příkazů ve formě syntaxe

[44] Zkuste si tento příkaz provést: vkopírujte jej do okna pro syntax a klikněte na ikonu zeleného trojúhelníku (nebo šipky, chcete-li).

Pozor, tečka na konci příkazu je bytostně důležitá. Pokud ji zapomenete, program neví, kde jeden příkaz končí a druhý začíná, takže výpočet odmítne.[45] Studenti, kteří s programem teprve začínají, příkazy obvykle zadávají klikáním na příslušné ikonky v rozbalovacích menu. Upozornění na možnost zadávat příkazy prostřednictvím syntaxe je především pro pokročilejší uživatele (více je o syntaxi v dodatku II této učebnice).

Pokud si v *Edit – Option – Viewer* (tedy v příkazech, jimiž nastavujeme vnitřní prostředí SPSS) zatržením kolonky u *Display command in log* (v levém dolním rohu interaktivního okna) tuto funkci nastavíte, máte možnost si po každém výpočtu zadaném v menu na prvních řádcích výsledků ve výstupu *Output* přečíst i text příkazu výpočet zadávající.[46]

Nová okna *Syntax* a *Script* lze otevřít v menu *File – New* a do otevřených oken lze psát konvenčním jazykem SPSS příkazy. Obsah okna lze uložit jako soubor syntaxí s příponou .sps a skriptů s příponou .sbs (jde o běžné textové soubory čitelné ve všech textových editorech). Soubory se syntaxí obsahují příkazy, které umožňují zadávat a spouštět statistické procedury (které jsou jinak v menu *Analyze*) a příkazy k transformaci dat (které jsou jinak v menu *Transform*). Skriptové soubory dovolují manipulovat s výstupy (oba typy souboru lze pro práci s daty kombinovat).

Prosíme čtenáře, aby se v tuto chvíli neděsili a pokračovali dále ve čtení. Tato nyní naprosto nepochopitelná hatmatilka se vám totiž po několika sezeních nad SPSS a práci s ním natolik dostane do krve, že se stane běžnou součástí vašeho datově analytického žargonu.

S psanými příkazy většinou nepracujeme, existují však užitečné výjimky. Zmíníme tři z nich:

- Syntax je výhodné použít při transformaci existujících proměnných do nové proměnné za pomoci logických podmínek – viz příslušnou kapitolu o transformaci proměnných a proceduře *If*.
- Je výhodné zapsat si syntakticky zadání rutinně opakovaného výpočtu s různými daty. Například tehdy, když se zabýváte problematikou nezaměstnanosti a úplně stejným způsobem zpracováváte začátkem každého měsíce data, která vám přicházejí ze statistického výkazu úřadu práce o počtech a struktuře nezaměstnaných. Jednou napsaný příkaz (syntax) slouží tak dlouho, jak zůstává výpočet neměnný. Pak stačí, abyste si otevřeli matici s novými daty a na ni pustili syntax uloženou na disku vašeho počítače prostřednictvím příkazu *Run*.

[45] Novější verze jsou ovšem již natolik „inteligentní", že se analytika zeptají, zda náhodou tečku nezapomněl.

[46] Pokud chceme zobrazit jen příkaz pro konkrétní operaci, pak po jejím naklikání přes menu stiskneme místo *OK* volbu *Paste*. SPSS operaci neprovede, pouze zobrazí příkaz do okna pro syntax. Pokud budeme chtít operaci z okna spustit, stačí tento příkaz označit myší a stisknout Ctrl + R. Můžete se takto alespoň částečně s příkazy naučit pracovat. Více se o příkazovém jazyce dozvíte ve druhém dodatku učebnice.

- U složitějších výpočtů vícerozměrných analýz je potřeba, abyste si všechny příkazy k analýzám uchovávali ve svém výpočetním archivu. Nikdy totiž nevíte, kdy si budete muset ověřit, zdali jste postupovali správně – a bez archivace syntaxe výpočtu toho nebudete schopni.

SPSS má poměrně rozsáhlou a dobře zpracovanou nápovědu (*Help*), která obsahuje i základní uvedení do programu (*Tutorial*). Rozhodně stojí za prohlédnutí. Společně s ukázkovými daty se totiž lze mnohému naučit sám i bez dotěrného učitele a (špatně) napsané učebnice.

2.2 Data

2.2.1 Matice dat

Při statistické analýze dat pracujeme s číslicemi, které mají určitý význam (pracujeme s kategorizovanými daty), nebo s čísly. Abychom mohli analýzu provádět, musíme tato data dostat do počítače a vytvořit v něm datovou matici. Protože jde o zpracování hromadných dat, pracujeme s hodnotami proměnných neboli s kvantifikovanými charakteristikami/vlastnostmi případů – to je respondentů či jiných objektů, popřípadě jevů. Matici tvoří tedy **případy** (obvykle řádky matice) versus **proměnné** (obvykle sloupce matice) a obsah matice tvoří **hodnoty příslušných proměnných** charakterizujících jednotlivé případy.

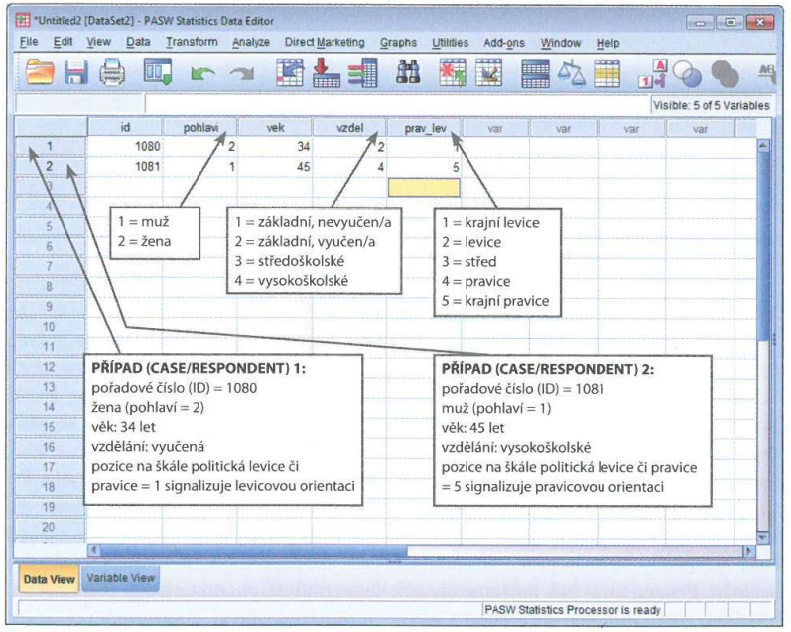
Případy jsou popsány svými vlastnostmi (atributy) – variantami neboli hodnotami proměnných, které jsou jejich logickými uskupeními. Například proměnná vzdělání může být uskupením možných nejvyšších dosažených stupňů vzdělání: základní, středoškolské, vysokoškolské, které lze popřípadě dále členit: vysokoškolské nižšího typu (Bc.), vyššího typu (Mgr.), popř. s vědeckou hodností (PhDr., Ph.D. apod.). Každému případu přidělujeme v matici jeho identifikační číslo – ID – a ideální je, máme-li stejným číslem označený i originální dokument, z něhož data o případu (nejčastěji respondentovi/respondentce – tedy dotazníku) čerpáme. Jen tak můžeme v případě nejasností porovnat zdroj dat s jejich záznamem v matici (proto dotazníky nikdy neničíme, ale archivujeme je) a data tak dodatečně kontrolovat. A že se chyby při nahrávání dat vyskytují, je mnohokrát potvrzenou zkušeností.[47]

Co jsou **proměnné**, již víme, stejně jako víme, že existují proměnné kategorizované (nominální a ordinální) a proměnné spojité (kardinální). U nominálních proměnných je spojení číslice (numerického kódu) a vlastnosti zcela arbitrární, takže bychom proměnnou „rodinný stav" mohli kódovat např. 1 = svobodný/á, 5 = ženatý/vdaná, 6 = rozvedený/á a 9 = ovdovělý/á, u ordinálních proměnných číslice označují pozici

[47] Což jenom dále nahrává určité skepsi o možnostech měření v sociálních vědách. Je ale pravda, že chybám při nahrávání dat do počítače se nevyhnou ani přírodní vědy.

varianty na škále, aniž by cokoliv říkaly o vzdálenosti mezi těmito pozicemi (vzdělání: základní = 1, střední = 2, vysokoškolské = 3 apod.). Je dobré si toto zvolené přiřazení číslic k charakteristikám pamatovat: je sice již dokumentováno v dotazníku, ale do něho nemůžeme při analýze z časových důvodů neustále nahlížet, a proto při definici matice v SPSS musíme vedle definice proměnných také určit, jak proměnnou pojmenujeme, kolik bude mít – v případě spojitých proměnných – desetinných míst, jaký verbální význam má jméno proměnné a číslice jejích hodnot v případě kategorizovaných proměnných. U spojitých proměnných jsou jejich hodnoty konkrétními čísly, která přímo vyjadřují množství příslušné vlastnosti, takže popisky jejich hodnot nejsou potřebné.

Zopakujme si tedy: každý případ představuje vektor obsahující **hodnoty příslušných proměnných** (každá varianta každé proměnné má přiřazenu číslici).[48] Vektory plníme do matice: co řádek, to případ (např. respondent), a co sloupec, to proměnná. Vše ilustruje obr. 2.6.



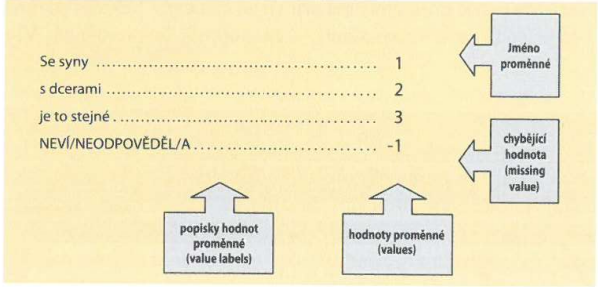
Obr. 2.6 Ukázka matice dat s pěti proměnnými a dvěma případy (respondenty) – smyšlený soubor

[48] SPSS umožňuje i záznam hodnoty proměnné ve formě textu. S ohledem na to, že tuto variantu používáme v sociálních vědách jen okrajově, z našich úvah ji vynecháváme.

Matice dat je, řečeno jinak, souborem kódů, za nimiž se skrývají konkrétní kvalitativní nebo kvantitativní vlastnosti jednotek našeho výzkumu. Tato data mohou být dále upravována, například pomocí transformací proměnných či výběrem případů a – především – analyzována.

Podíváme-li se nyní na dotazník, jenž je v sociologii obvyklým nástrojem sběru kvantitativních údajů, z hlediska analýzy dat, je nutno každou jeho otázku (položku) a navržený způsob měření (to je k ní přiřazenou stupnici hodnot) vnímat tak, jak ukazuje obr. 2.7.

36. S kým jsou podle Vás větší výdaje, se syny nebo s dcerami anebo je to stejné?



Obr. 2.7 Demonstrace otázky v dotazníku jako (kategorizované) proměnné z hlediska jazyka SPSS

Pozn. Odpovídající jméno proměnné se obvykle neuvádí v dotazníku (proto zde absentuje), ale až při vytváření matice dat.

Každé otázce v dotazníku (neboli proměnné) musíme přiřadit technické jméno: v matici dat se toto jméno objeví v označení sloupce. Jména (*names*) můžeme dávat různá, ale SPSS má svá přísná pravidla, jak mohou vypadat. Především, každé jméno musí začínat písmenem, jména by ale měla být přiměřeně krátká.[49] Ve jménech můžeme používat písmena i číslice, jména tedy mohou být alfanumerická, ale bez diakritiky (např. *P1*, *A26*, *X26*), lze používat i podtržítko (např. *P_01*). Při vymýšlení jmen si můžeme stanovit systém, který nám usnadní práci s maticí při analýze. Například rezervujeme všechna jména začínající písmenem A pro proměnné, které zjišťovaly důvody, proč lidé chtějí mít děti, písmeno B pro důvody, proč děti mít nechtějí. Písmenem D pak můžeme označit demografické charakteristiky respondentů (pohlaví, věk, rodinný stav apod.) Nebo můžeme používat i nápovědných jmen proměnných: např. *sex* pro pohlaví (ne pro frekvenci koitu), *prijem*, *vzdel*, *job* (povolání) apod. Abychom věděli přesně, co tato technická jména znamenají, SPSS

[49] Starší verze SPSS umožňovaly užít jen 8 znaků. Je vhodné tento limit nepřekračovat, protože jiné pakety mají tento limit dodneška.

umožňuje jména také popsat ve formě *variable label* (viz obr. 2.3) – v podstatě sem můžeme vložit zkrácenou verzi otázky.

Pro popisky variant odpovědí (*value labels*) už můžeme využít diakritických znamének (možnost psát česky si lze nastavit v menu *Edit – Options*). Doporučujeme, aby tyto popisky byly přiměřené krátké (ale výstižné), neboť se zobrazují ve výstupních tabulkách a je jasné, že každá tabulka má při tisku na ploše stránky formátu A4 jen omezený prostor. Jednotlivým *value labels* odpovídají příslušné hodnoty (*values*), v našem příkladu to jsou hodnoty 1, 2 a 3. O chybějící hodnotě (*missing value*) se zmiňujeme níže.

V některých případech mají otázky v dotazníku podobu tzv. baterie otázek, což ilustruje obr. 2.8.

1. Řekněte prosím o každé z následujících skutečností, jak je ve Vašem životě důležitá:

	Velmi důležitá	Dost důležitá	Ne příliš důležitá	Vůbec ne důležitá	Neví	Neodpověděl(a)	
A Práce	1	2	3	4	–1	–2	Q1a
B Rodina	1	2	3	4	–1	–2	Q1b
C Přátelé a známí	1	2	3	4	–1	–2	Q1c
D Volný čas	1	2	3	4	–1	–2	Q1d
E Politika	1	2	3	4	–1	–2	Q1a
F Náboženství	1	2	3	4	–1	–2	Q1f

Obr. 2.8 Demonstrace baterie otázek v dotazníku jako sady (kategorizovaných) proměnných

V této tabulce je každý řádek proměnnou s variantami/oborem hodnot <1; 4>; zde jsme zvolili názvy *Q1a* až *Q1f* (mohlo by být také *Q1_1* až *Q1_6*) proto, aby napovídaly, že všech 6 proměnných tvoří jednu společnou baterii otázek, že tedy měří něco společného. Záporné hodnoty u odpovědí „neví" a „neodpověděl(a)" představují svým způsobem chybějící údaje neboli **chybějící hodnoty** (*missing values*) – vždycky musíme totiž předpokládat, že někteří respondenti na naše otázky neodpoví. I chybějící údaje je dobré kódovat, volba kódu je libovolná, kód však nesmí mít hodnotu, kterou může nabývat příslušná proměnná. U *missing values* je navíc někdy užitečné rozlišovat, kdy údaj chybí proto, že respondent na otázku odmítl odpovědět, kdy proto, že se ho netýká (což je případ v naší tabulce), popřípadě že chybějící údaj vznikl opomenutím tazatele apod.[50]

[50] V některých situacích má totiž smysl analyzovat i tyto chybějící odpovědi, zvláště když se jich u některých položek objeví mnoho: analýza nám umožní zjistit, jaký typ respondentů odpověď odmítl, což může být pro intepretaci výsledků důležitá informace.

2.2.2 Definice jednotlivých proměnných

Abychom mohli matici naplnit daty, musíme ji, jak jsme naznačili výše, nejprve definovat. Děje se tak v okně *Variable View*, které jsme ukázali již na obr. 2.3. Jde o tyto úkony:

- Připsání technického jména proměnné, určení jejího místa v matici (sloupce/sloupců).
- Definice charakteru proměnné jako **číselné** (*numeric*) či **textové** (*string*); textovou proměnnou počítač chápe jako označení a neprovádí s ní početní operace, jako je sčítání či násobení.
- Připsání širšího/podrobnějšího **označení proměnné** (*variable labels*).
- Připsání verbálního **označení jednotlivým hodnotám** (kategorizované) proměnné (*value labels*). Ty zpřehledňují tištěné výstupy, neboť přiřazují ke jménům proměnných i vysvětlující popis. Např. q1_2 může být jméno proměnné neboli *variable name* a „důležitost rodiny v životě" může být vysvětlující popisek jména této proměnné neboli *variable label*. Její varianty „velmi důležitá", „dosti důležitá", „nepříliš důležitá", „vůbec nedůležitá" a „neví" jsou pak *value labels* dané proměnné.

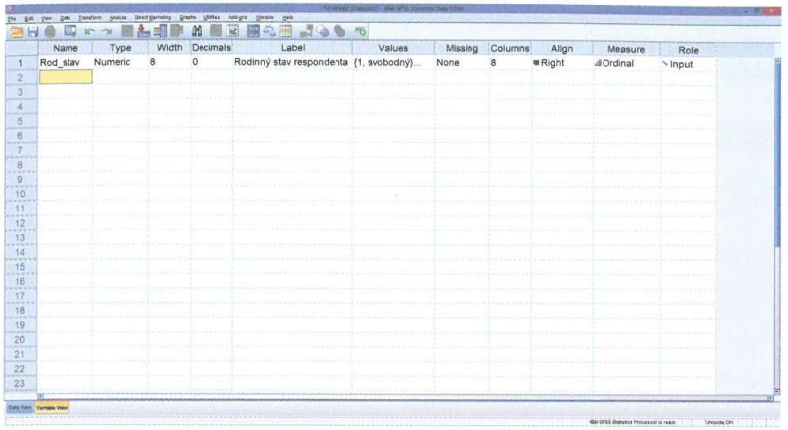
Pozor: V případě spojitých proměnných nedávají *value labels* smysl a nepoužíváme je!

- Určení počtu desetinných míst (v případě spojitých proměnných).
- Definování **uživatelských chybějících hodnot** (tzv. *user missing values*).[51] Někdy do *missing value* přeřazujeme některé hodnoty proměnných při jejich transformaci. Týká se to například varianty „nevím", která sice někdy může být součástí ordinální proměnné jakožto její středová hodnota (1 = s vládou jsem spokojen, 2 = nevím, 3 = s vládou jsem nespokojen), častější jsou ale případy, kdy ji používáme jen proto, abychom nenutili respondenta/respondentku zaujímat postoj, který nemá. V další analýze se pak často soustřeďujeme jen na ty, kdo postoj zaujali a v modulu *Transform – Recode* přiřadíme odpovědím „nevím" číslici označující chybějící hodnotu (*missing value*). Často je to záporná hodnota.[52]
- Rozumné je určit u každé proměnné v kolonce *Measure* úroveň měření. Tato informace totiž u některých statistických operací rozhoduje o volbě počítaných statistik.

Některé sloupce v okně *Variable view* (viz obr. 2.9) můžeme vyplnit přímo vepsáním příslušného textu, jiné nám nabídnou po kliknutí předdefinované volby. Přímo vyplňujeme sloupce *Name* a *Label*, v ostatních po kliknutí vyskočí rozbalovací okno.

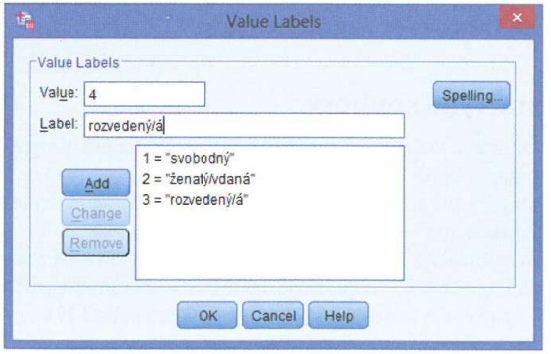
[51] SPSS zná i **systémové chybějící hodnoty** (*system missing value*). Ty se zobrazují v datové matici jako tečky a znamenají, že pro danou proměnnou a daný případ není k dispozici žádná hodnota (např. v dotazníku není zaškrtnuta žádná odpověď).

[52] K tomu, jak se v některých konkrétních statistických technikách zachází s *missing value*, se ještě vrátíme. Většinou se s případy obsahujícími *missing value* nepracuje (s výjimkou uvedenou v pozn. 48).



Obr. 2.9 Okno pro popis proměnných

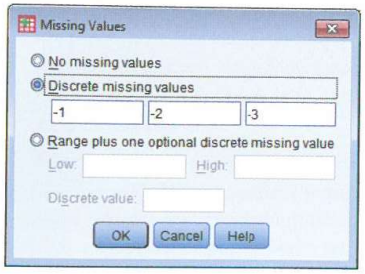
Klikneme-li na sloupec *Values*, můžeme v rozbalivším se okně vepsat popisky jejich hodnot (viz obr. 2.10). Do kolonky *Value* vepíšeme číslici hodnoty, tabulátorem či pomocí myši přejdeme do kolonky *Label* a vepíšeme popisek. Spodní tlačítka nám umožní takto definovaný *label* přidat (volba *Add*) a v seznamu *labels* pak provádět změny (volby *Change* či *Remove*). Nakonec vše odsouhlasíme kliknutím na *OK*.



Obr. 2.10 Okno pro popis variant znaků (*Value Labels*)

Podobně můžeme definovat i uživatelské chybějící hodnoty (*missing values*), což jsou hodnoty, které nevcházejí (pokud si to výslovně nepřejeme a nezadáme příkazem) do analýzy. SPSS nám k tomu nabízí speciální okno, jehož ukázku uvádí obr. 2.11.

Zde jsme rozhodli, že pro chybějící hodnoty budeme rezervovat výrazy −1, −2 a −3. Jistě podle ukázky přijdete sami na to, jaké další možnosti, jak definovat *missing values*, se nabízejí.[53]



Obr. 2.11 Okno pro definování chybějících hodnot

2.2.3 Plnění matice dat

Data můžeme dostat do matice různými způsoby. Důležité jsou pro nás zejména:

- Plnění námi definované matice našimi daty.
- Import dat ze souboru jiného typu (z textového editoru, databáze či tabulkového procesoru – *spreadsheetu* –, např. programu, jako je Excel).

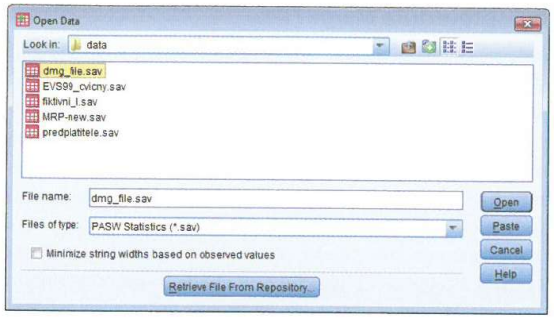
Můžeme ovšem také použít i dříve nebo někým jiným vytvořenou matici dat (tzv. systémový soubor). Nejčastějším případem je plnění námi definované matice dat přepisem údajů z dotazníků nebo jiných záznamových archů.

2.3 Práce se systémovými soubory

Existující datové soubory otvíráme stejně jako jakékoliv soubory v jiných programech. Tedy po spuštění programu SPSS klikáme postupně na tlačítka *File – Open – Data*, v otevřeném okně pak najdeme to správné místo na disku (popř. na externím disku nebo flash disku), kde máme soubor uložen (viz obr. 2.12).

Máme-li matici naplněnou našimi daty, snažíme se tuto matici zachovat pro další zpracování tím, že ji uložíme jako systémový soubor. SPSS takovým souborům při jejich uložení přidává příponu .sav – podle ní tyto soubory můžete identifikovat (obr. 2.13).

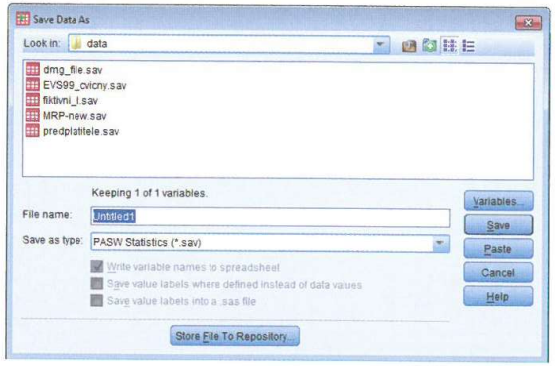
[53] SPSS je v možnosti definovat *missing values* nesmírně flexibilní, jiné statistické balíky umožňují zpravidla definovat pouze jedinou hodnotu jako chybějící.



Obr. 2.12 Okno pro otvírání souboru uloženého na disku nebo externím nosiči

Soubor ukládejte průběžně, to je v každém kroku popisu a plnění matice stejně jako po každé změně, kterou v ní provedete (např. po přidání případu nebo vytvoření nových proměnných – viz kapitolu 6, věnovanou transformaci proměnných). Ponechávejte přitom (samozřejmě pod různými názvy):

- pramenný soubor, což je naplněná a zkontrolovaná původní matice, v níž nebyly provedeny žádné další změny;
- předposlední podobu souboru (po předposledních provedených změnách);
- poslední podobu souboru (po posledních provedených změnách).



Obr. 2.13 Ukládání dat

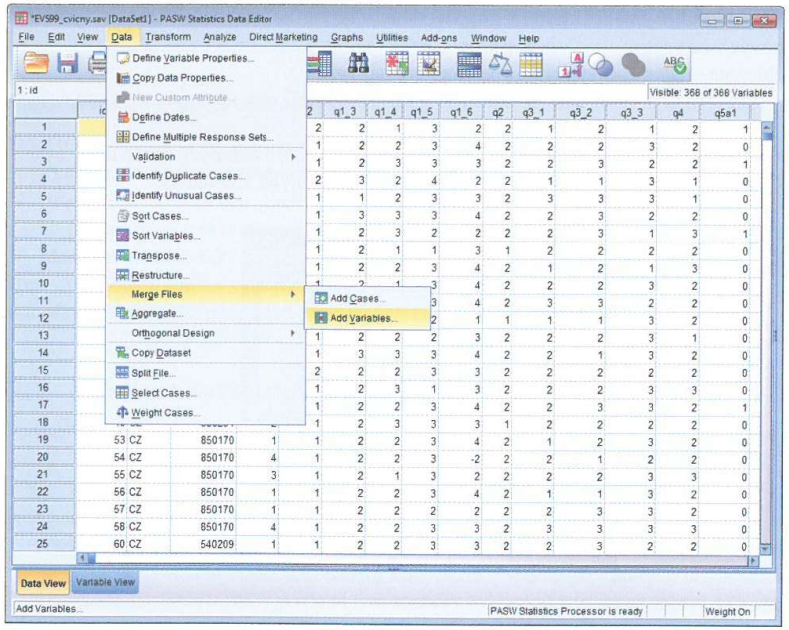
Plnění těchto zásad se vám vyplatí! Někdy se mohou totiž naplnit i katastrofické scénáře a při práci s poslední verzí souboru o něj můžete v důsledku technických potíží programu nebo počítače přijít. Uchovávejte proto raději starší verze souboru i mimo harddisk svého počítače. Nebo se může stát, že omylem provedete při transformaci proměnných v matici nevratné změny, jak ukážeme v lekci o transformaci proměnných.

2.3.1 Slučování souborů (procedura *Merge Files*)

V některých případech potřebujeme k analýze data, která se nacházejí v různých souborech. Tyto soubory je nutné sloučit. Lze tak učinit několika způsoby:

a) Procedura *Add Variables* (přidání dalších proměnných)

Máme v jedné databázi (matici) údaje o osobních charakteristikách studentů a v druhé databázi (matici) údaje o jejich prospěchu. Chceme je dostat do jedné matice, abychom měli o studentech všechny údaje pohromadě. Aby to bylo možné, musí být pořadí studentů v obou maticích shodné (to aby se příslušné údaje připisovaly příslušnému studentovi), nebo musíme mít nějaký znak, který každého studenta jednoznačně definuje (nejlépe ID). Při operaci přidávání proměnných se k proměnným jednoho souboru přidají proměnné dalšího souboru, jak naznačuje následující schéma (viz obr. 2.14 a schéma 2.1).



Obr. 2.14 Přidávání dalších proměnných k existujícím datům

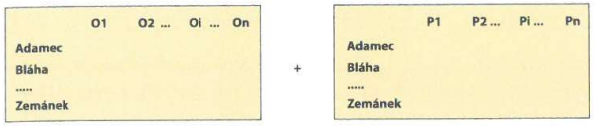
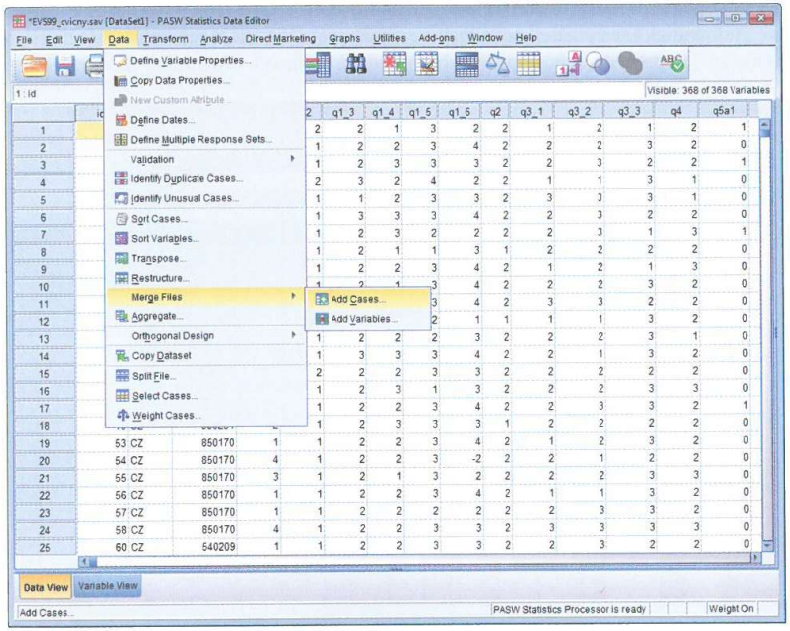


Schéma 2.1

b) Procedura *Add Cases* (přidání dalších případů)

Jindy k existujícím případům potřebujeme dohrát jen několik dalších. K tomu slouží procedura *Add Cases* (viz obr. 2.15 a schéma 2.2). Tato situace vzniká např. tehdy, když máme personální databáze jednotlivých imatrikulačních ročníků studentů (každý ročník je samostatná matice dat) a chceme vytvořit jednotnou databázi studentů všech ročníků (jednu matici). Struktura matice je stejná: sledují se stejné proměnné (charakteristiky studentů) a v maticích jsou uvedeny ve stejném pořadí. K případům jednoho souboru se přidají případy druhého souboru.



Obr. 2.15 Přidávání nových případů k již existující matici

	A1	A2	A3	A4	A5	A6	.....	Ai	An
Adamec									
Bláha									
.....									
Zemánek									

+

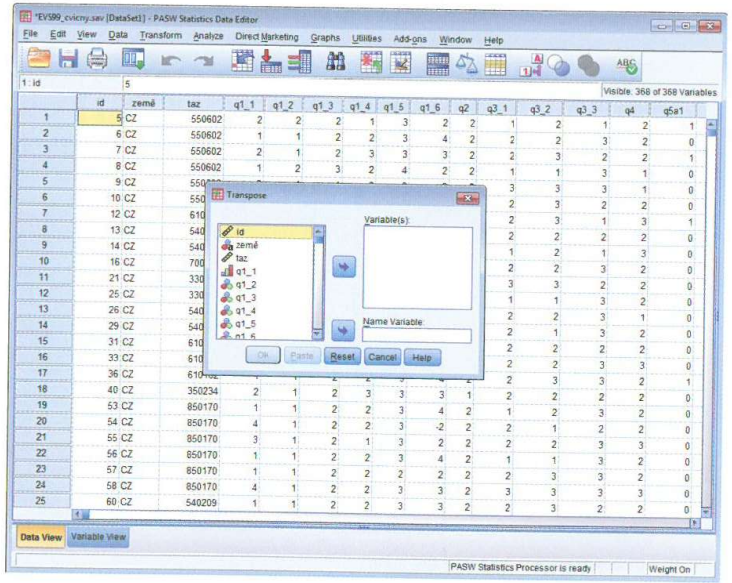
	A1	A2	A3	A4	A5	A6	.....	Ai	An
Deml									
Stará									
.....									
Vechtr									

Schéma 2.2

2.3.2 Záměna řádků a sloupců matice (procedura *Transponse*)

Příkaz *Transponse* vytváří nový datový soubor, ve kterém jsou:
- původní řádky (případy) sloupci (proměnnými);
- původní sloupce (proměnné) řádky (případy).

Automaticky se vytvářejí nová jména proměnných. Využití je zejména ve složitějších statistických procedurách, případně při převodech do jiných programů (viz obr. 2.16).



Obr. 2.16 Vzájemná záměna řádků a sloupců matice

2.4 Výběr případů z výběrového souboru

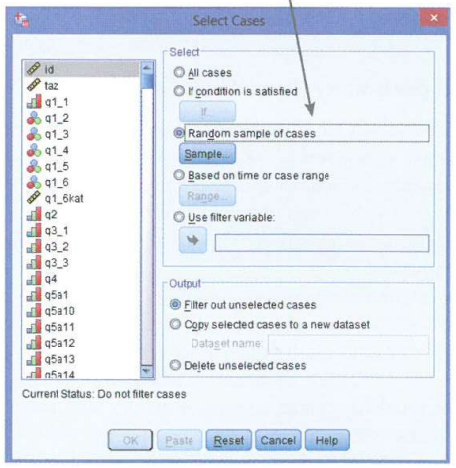
Výběr případů představuje manipulaci s datovým souborem, která nám umožní pracovat pouze s určitým podsouborem případů. Pomocí procedury *Data – Select Cases* můžeme požadovaný podsoubor definovat:

a) Podsoubor náhodně vybraných případů, máme-li například příliš velký soubor, jako tomu může být v případě dat z mikrocensu apod. Důvod redukce velikosti našeho výběrového souboru může být v tomto případě technický – operace probíhají rychleji. Navíc, kupodivu, mohou být naše výsledky přesnější.

b) Podsoubor vybraný na základě výzkumné otázky. Chceme provádět výpočty například jen s lidmi, kteří preferují určitou politickou stranu, nebo s lidmi určité věkové skupiny (například s osobami staršími 60 let) apod. Pozor ale: práce s podsoubory předpokládá, že výběrový soubor je natolik velký, že má statistický smysl z něj vybírat soubor menší, podsoubor.

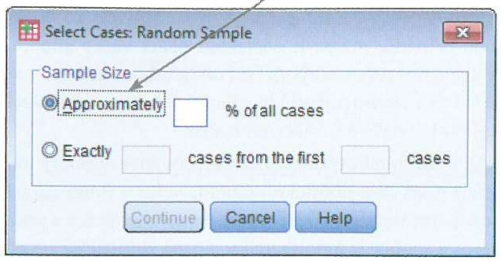
2.4.1 Výběr případů prostřednictvím pravděpodobnostního (náhodného) výběru (procedura *Random sample of cases*)

Operace *Random sample of cases* dovoluje vytvořit z našeho pracovního (výběrového) souboru pravděpodobnostní náhodný výběr tím, že omezíme počet jeho jednotek. Pokud byl náš původní soubor reprezentativní, bude při tomto způsobu výběru i náš nově vybraný (pod)soubor reprezentativní, viz obr. 2.17.



Obr. 2.17 Výběr náhodného podsouboru z původního souboru

Můžeme buď vybrat přibližný (*Approximately*) podíl z původního souboru (např. 25 %), nebo určitý počet případů (po kliknutí na možnost *Exactly* vypíšeme celkový počet jednotek původního souboru).



Můžeme zadat procento případů z celku, které chceme do výběru zahrnout, nebo přesný počet případů.

Co se týče rozhodnutí, co s nevybranými případy (viz možnosti v rámečku *Output* na obr. 2.17), doporučujeme používat raději variantu *Filter out unselected cases* (na obrázku je zapnuta), která nevybrané případy nemaže, ponechává je v souboru, ale nepracuje s nimi (v matici dat takové případy poznáme podle toho, že jejich ID je přeškrtnuto). Filtr lze totiž lehce vypnout, takže pokud je potřeba (a ono to většinou potřeba je), lze dále pracovat s celým souborem. Pokud použijeme variantu *Delete unselected cases*, jsou všechny nevybrané případy odstraněny. Pak ale musíme být velmi opatrní a přemýšlet, zdali takto redukovaný soubor chceme uložit, nebo ne. Pokud ho uložíme pod stejným jménem, původní soubor se přepíše a nám zůstane jen soubor s vybranými jednotkami – a právem také oči pro pláč, pokud bychom neměli předposlední podobu souboru zálohovanou.

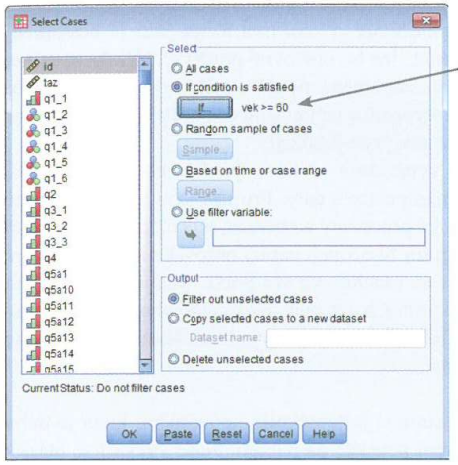
2.4.2 Výběr případů za pomoci podmínky (procedura *Select cases if*)

Někdy se může stát, že nás analyticky zajímají jen menší podsoubory (například jen ženy nebo jen osoby se středoškolským vzděláním a vyšším, popřípadě jen osoby bydlící v Praze), a proto si je vybíráme, abychom další analytické výpočty prováděli jen s nimi. Je pochopitelné, že je můžeme vybírat jen podle známých – zjištěných – charakteristik: pokud jsme například v dotazníku nezjišťovali místo bydliště respondenta, nemůžeme vybrat, řekněme, obyvatele Prahy.

Podsoubory, s nimiž chceme pracovat, určujeme pomocí podmínky: do okénka vyklikáme nebo vypíšeme podmínku, např. SEX = 1 (chceme-li pracovat jen s muži a víme, že v proměnné SEX 1 = muž), OBEC = 15 (chceme-li pracovat jen s obyvateli Prahy a víme, že v proměnné OBEC Praha = 15), VZDEL > 2 (chceme-li pracovat s osobami, jež mají středoškolské a vysokoškolské vzdělání, a víme, že v proměnné VZDEL osoba se středoškolským vzděláním = 3 a osoba s vysokoškolským vzděláním = 4).

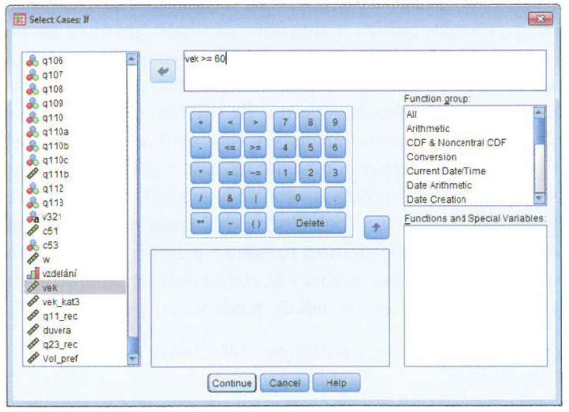
Může nás například zajímat analýza lidí ve věku 60 let a starších. K výběru takového podsouboru použijeme proceduru *Data – Select cases – If condition is satisfied*.

Po kliknutí na tlačítko *If...* se objeví dialogové okno, do nějž vepíšeme příslušnou podmínku pro výběr (viz obr. 2.18 a 2.19). V našem případě vybíráme podsoubor respondentů ve věku 60 let a starších.



Toho dosáhneme, klikneme-li na *If* a zadáme-li podmínku výběru.

Obr. 2.18 Způsob výběru podsouboru jednotek



Obr. 2.19 Zadávání podmínek pro výběr podsouboru

Podmínky lze samozřejmě různě kombinovat, např. by bylo možné získat jen podsoubor mužů ve věku 60+ let, kteří ještě pracují, apod. Přidání dalších podmínek se řídí pravidly logiky a my si přitom musíme dávat dobrý pozor, co z hlediska výběru podsouboru znamenají příkazy „nebo" a „a současně" (blíže o tomto detailu v 6. kapitole).

Před přechodem k výpočtům opět s celým souborem nesmíme zapomenout filtraci odstranit. Chceme-li v průběhu práce s daty ukončit práci s vybraným podsouborem a vrátit se k celému souboru, klikneme v *Select cases* na *Reset* (filtr je odstraněn) nebo na *All cases* (filtr je pouze vypnut, lze ho opětovně použít).[54] Někdy se stane, zejména v časové tísni, že na to člověk zapomene, používá stále vybraný podsoubor a výsledky mylně vydává za produkt výpočtů s celým souborem. Pak je samozřejmě překvapen, jaké „neuvěřitelné" věci z analýzy vycházejí.[55]

Všechny výše uvedené postupy využíváme většinou před samotnou analýzou, takže mají spíše povahu technické manipulace s daty. Pro nás jsou ale samozřejmě mnohem důležitější příslušné statistické procedury a operace, jejichž prostřednictvím získáváme výzkumné výsledky. Program SPSS jich nabízí obrovské množství, které sociolog málokdy detailně zvládne a ne všechny ve své praxi využije – ať je jejich využití limitováno povahou jeho výzkumných dat, nebo rozsahem jeho znalostí statistiky. S některými z nich – s těmi nejfrekventovanějšími a jednoduššími – se postupně seznámíme v následujících kapitolách.

Než se do toho pustíme, neodpustíme si ještě několik upozornění, která je třeba mít při práci se statistickým softwarem neustále na paměti: Naše výzkumné otázky (a tedy i charakter získávaných dat) bychom nikdy neměli přizpůsobovat statistickým procedurám, které nabízí náš program, ale naopak bychom měli vyhledávat procedury umožňující maximální využití našich dat. To ale vyžaduje, abychom 1) měli alespoň povědomí o tom, co každá z procedur nabízí, a abychom 2) věděli, jaké má každá procedura požadavky na povahu škál, s nimiž pracuje (tedy jaké jsou podmínky její aplikace). To neruší požadavek důrazu na věcnou stránku výzkumu, upozorňuje nás to však na to, že základní plán analýzy dat musíme mít již při koncipování výzkumu. Abychom mohli použít například faktorovou analýzu (její aplikace ovšem musí vycházet z toho, že nám pomůže smysluplně odpovědět na naše výzkumné otázky, a nikoliv z toho, že jsme se ji právě naučili používat nebo že je módní), musíme již ve svém dotazníku pro ni připravit vhodné otázky (položky): tedy že musíme formulovat baterii alespoň 6–10 otázek se stupnicemi o stejném rozsahu s alespoň 5 hodnotami. 3) Musíme mít neustále na paměti, že program spočítá vše, co mu zadáme, a nepřemýšlí (ani nemůže, není to myslící bytost, byť máme někdy tendenci jej personifikovat)

[54] Pokud chcete začít pracovat s celým souborem a použili jste volbu *Delete unseelcted cases*, musíte si původní soubor znovu otevřít!

[55] Na tuto chybu občas narážíme v bakalářských pracích. I to je jeden z důvodů, proč po studentech požadujeme, aby s finálním textem práce odevzdávali i datový soubor. Nezdají-li se oponentům některé výsledky, velmi rychle si je zkontrolují.

o smysluplnosti zadání, ani o dodržení určitých požadavků na analýzu. My musíme například vědět (ne ON), že nemá příliš smysl sledovat souvislost dvou proměnných pouze prostřednictvím charakteristiky chí-kvadrát (vysvětlíme v kapitole 8) nebo že není možné počítat z kódů nominální proměnné průměr a že je nemyslitelné, abychom při aplikaci lineární regrese měli naši závisle proměnnou dichotomické povahy. Pokud si některá základní pravidla statistiky nebudeme pamatovat, může se stát, že budeme často počítat (s přesností několika desetinných míst) a pak i sofistikovaně interpretovat naprosté nesmysly. Eufemisticky, abychom neurazili, tomu říkáme „produkovat statistické artefakty".[56]

[56] Angličtina pro to má pěkný akronym GIGO = *garbage in, garbage out* (smetí vevnitř, smetí venku).