Kapitola 3

Základy jednorozměrné analýzy

Jednorozměrná analýza je základním vhledem do datového souboru. Je popisem souboru z hlediska jedné proměnné a představuje základ deskriptivního výzkumu, tedy popis toho, jak se věci mají, bez ambicí vysvětlovat, proč tomu tak je. Jelikož analyzujeme, to je třídíme soubor jednotek pouze z hlediska jedné proměnné, hovoříme také o třídění I. stupně. Prozkoumání (exploraci) proměnné děláme tak, že se 1) díváme, jak je proměnná rozložena, to znamená jakých nabývá hodnot a jak často jsou tyto hodnoty zastoupeny; abychom tomuto rozložení lépe porozuměli, doporučujeme začít analýzu prostřednictvím grafu, který nám lépe než tabelovaná řada čísel dá představu o distribuci hodnot dané proměnné a o tvaru tohoto rozložení;[57] 2) vypočítáváme příslušné souhrnné statistiky (střední hodnoty), které umožňují prostřednictvím jednoho čísla vystihnout základní charakteristiku zkoumané proměnné. Tomuto typu analýzy se také říká **explorační** (vyhledávací) **analýza**.

V samotném počátku práce s daty, především s těmi novými, která jsme získali naším vlastním výzkumem, se jednorozměrná analýza také používá pro čištění dat (to je pro odstraňování chyb) a pro úvahy, jak proměnné transformovat.

Prvním krokem, který musíme udělat před jakoukoliv analýzou dat, je tzv. **čištění dat**. Nejedná se o nic jiného než o kontrolu dat – to je zdali při jejich nahrávání[58] nedošlo k chybě, zdali jsme nenahráli jiné hodnoty, než které jsme zjistili ve výzkumu. Zejména jsou odhalovány hodnoty mimo povolený obor hodnot, například když bylo omylem kódováno pohlaví jedince číslicí 3, ačkoliv povolený obor hodnot je <0; 1> (0 = muž, 1 = žena). Může jít také o podklad pro úpravy souboru, například odstranění odlehlých případů u spojitých proměnných, když u některých jednotek jsme zjistili hodnoty, které jsou velmi atypické (kupříkladu extrémně vysoké údaje o příjmu některého z respondentů, které mohou vypadat „podezřele" a jsou nepravděpodobné).

---

[57] V analýze dat často uslyšíte okřídlené rčení, že graf je lepší než tisíc slov.

[58] Nahráváním rozumíme vyplnění datové matice (viz předchozí kapitola).

Čištění dat je důležité, protože některé analytické postupy jsou velmi citlivé na hodnoty, které jsou výrazně nižší nebo naopak výrazně vyšší, než je převážná většina hodnot dané proměnné. V jazyce datové analýzy se jim říká **odlehlé hodnoty** (*outliers*), popřípadě extrémně odlišné hodnoty. Odlehlé hodnoty často vznikají chybou při nahrávání: např. přidáním řádu, když při nahrávání měsíčního příjmu respondenta, který je 15 800 Kč, ve skutečnosti nahrajeme 158 000 apod. Pozor ale, odlehlé hodnoty nemusejí být vždy jen produkty našich chyb při vyplňování datové matice, mohou být údajem o reálné situaci a my pak stojíme před rozhodnutím, jak s nimi naložit. Například zjistíme, že hrubý měsíční příjem většiny jednotek v souboru se pohybuje mezi 9 000 až 40 000 Kč, avšak my máme v souboru dvě jednotky, z nichž jedna uvádí příjem 800 Kč a druhá 40 000 000 Kč. Můžeme je z analýzy odstranit, neboť tato čísla jsou buď pokusem o žert (pak jsou ale všechny hodnoty obou respondentů nedůvěryhodné), nebo to jsou sice reálné údaje, ale vzhledem k tomu, že jde o dva zcela atypické jedince, jejich údaje nám k ničemu nejsou, neboť jsme do výzkumu zachytili osamělé reprezentanty marginálních sociálních skupin či kategorií. Navíc ponecháním jedince se čtyřicetimilionovým příjmem výrazně zkreslíme údaj o průměrném příjmu souboru – dojde k jeho výraznému zvýšení. (Později si ukážeme, že je možné vypočítat i tzv. ořezaný průměr – *trimmed mean* –, který nás tohoto problému zbaví.)

Čištění dat není příliš záživná činnost (upřímně řečeno, je to dost velká nuda), nicméně je to činnost naprosto nezbytná. Žádný odpovědný badatel a analytik nezačne s vlastními analýzami dříve, pokud nemá jistotu, že má všechna data zkontrolována a vyčištěna. V hlavě mu totiž varovně bliká již zmíněný okřídlený akronym GIGO (*garbage in, garbage out, smetí dovnitř, smetí ven*), který upozorňuje, že pokud nahrajeme špatná data, je logické, že i výsledky našich analýz budou špatné. Jak připomíná Swoboda (1977), nesmíme zapomínat na to, že žádná statistika není lepší než její surovina. Tak jako nemůže být správný úsudek, nejsou-li správné předpoklady, stejně tak nejsou k ničemu i ty nejobtížnější početní operace, pokud číselný materiál je hned od počátku nesprávný nebo nedostačující. Swoboda dále poznamenává, že početní chyby lze napravit a nevhodné metody nahradit lepšími, ale máme-li chybné či nesprávné (irelevantní či nevalidní) prvotní údaje, pak přes nasazení nejsofistikovanějších technik jejich zpracování a přes sebevětší pečlivost v dalším postupu již jen množíme chyby.

Po vyčištění souboru lze zahájit vlastní analytické práce s daty. Ty začínáme vždy statistickým popisem neboli deskripcí struktury souboru. **Deskripce struktury souboru** podle jednotlivých charakteristik výzkumných jednotek spočívá ve zjištění rozložení (rozdělení, distribuce) četností hodnot proměnné. Zajímáme se o to, kolik máme jednotek s určitou vlastností: například žen orientovaných především na rodinu, žen orientovaných na práci a zaměstnání a žen smíšeného typu[59] a jaký podíl ve výběrovém souboru představují.[60] V případě spojitých znaků mohou být popisem

[59] Mimochodem, toto je typologie Catherine Hakim (2000) z její preferenční teorie.

[60] Na základě reprezentativního šetření provedeného v roce 2005 druhým z autorů této učebnice dnes víme, že mezi českými ženami bylo ve věku 20–40 let 17 % žen orientovaných na rodinu, 13 % žen orientovaných na zaměstnání a 70 % žen smíšeného typu.

jejich střední hodnoty, například průměrný věk nebo medián příjmového rozložení apod.

Úvahy o transformacích proměnných. Na základě znalosti rozložení jednotlivých proměnných a jejich statistických charakteristik je možné zvažovat, zda se proměnné nemají vhodným způsobem upravit (transformovat), aby vyhovovaly jednotlivým statistickým procedurám. Můžeme například slučovat několik kategorií proměnné, abychom měli dostatečně zastoupené kategorie (např. u kontingenčních tabulek – viz kapitolu 8), nebo různými matematickými operacemi měnit rozložení proměnné tak, aby lépe vyhovovalo statistické analýze (vše bude blíže vysvětleno v následujících kapitolách této učebnice).

3.1 Rozložení kategorizovaných dat

3.1.1 Čištění dat – jak na to

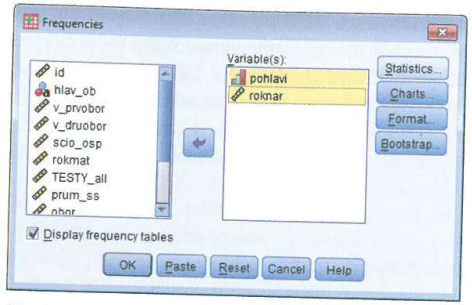
Kontrola chybných dat spočívá v tom, že pečlivě pozorujeme, zdali jednotlivé hodnoty variant znaku (proměnné) odpovídají variantám, které máme v dotazníku. Díváme se tedy, řečeno jinými slovy, zdali se distribuce (rozložení) nahraných hodnot pohybují pouze v rámci stupnic, s jejichž pomocí jsme jednotlivé proměnné měřili. Kontrolujeme samozřejmě všechny proměnné, které naše datová matice obsahuje, ale způsob kontroly závisí na typu proměnné. U kategorizovaných dat neboli u proměnných nominálních a ordinálních a také u proměnných intervalových s malým počtem variant (např. počet dětí respondenta) je způsob kontroly odlišný od způsobu kontroly spojitých (nekategorizovaných) proměnných, to je proměnných intervalových s velkým počtem variant (jimiž jsou např. věk, příjem, IQ skóre, skóre v přijímacím testu atd.).

Data kontrolujeme tak, že si necháme udělat rozložení četností jednotlivých proměnných. K tomu použijeme proceduru z SPSS *Analyze – Descriptive Statistics – Frequencies* a v rámci *Frequencies* si ještě necháme spočítat minimální a maximální hodnotu znaku – to pro kontrolu kardinálních proměnných s velkým počtem variant. Ukažme si vše na příkladu. Použijeme k tomu soubor dat o přijímacím řízení (viz soubor „fiktivni.sav“).[61] V tomto souboru zkontrolujeme proměnné pohlaví a rok narození (pojmenované jako „pohlavi“ a „roknar“). Rok narození bývá ve většině výzkumů proměnná, která má velký počet variant, takže bychom ji měli chápat jako proměnnou nekategorizovanou, ale v našem případě – jelikož se jedná o uchazeče o prezenční studium na VŠ – bude mít variant jenom omezený počet (dokážete říci proč?).[62] Je to tedy vhodná proměnná pro tuto proceduru. Nuže, jak postupujeme?

[61] Je to soubor vytvořený speciálně pro potřeby tohoto kurzu – z fiktivního přijímacího řízení v roce 1998, v němž byly záměrně vytvořeny chyby při nahrávání.

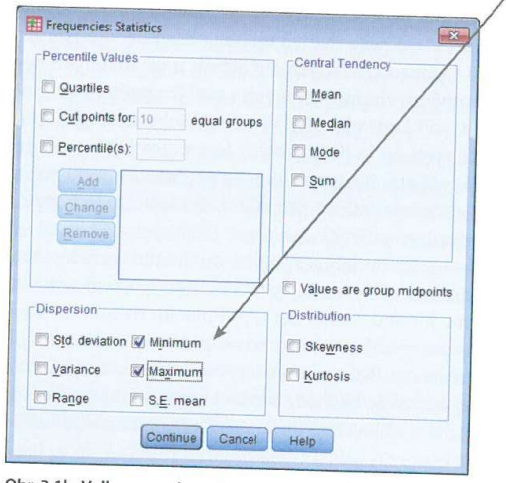
[62] Je to z toho důvodu, že na vysokou školu se hlásí především mladí lidé, kteří jsou si věkově podobní, a jsou tudíž narozeni víceméně ve stejném roce, v našem případě kolem roku 1980.

Ve *Frequencies* klikneme na jména těch proměnných, které chceme kontrolovat, a přesuneme je do okna *Variable(s)*. V našem příkladu to jsou to proměnné „pohlavi" a „roknar" (viz obr. 3.1a).[63]



Obr. 3.1a Způsob zadávání operace *Frequencies*

Klikneme na tlačítko *Statistics* – a zaškrtnutím příslušných políček zvolíme nalezení minimální a maximální hodnoty (obr. 3.1b).



Obr. 3.1b Volba pro nalezení minimální a maximální hodnoty zkoumaných proměnných

[63] Sluší se dodat, že po získání nových dat (vlastních či převzatých) je vhodné zkontrolovat rozpětí všech proměnných. V případě, že jsou proměnných stovky, zabere nám tato kontrola mnoho hodin nudné, ale potřebné práce. Kontrolu rozsahů a jiné kontroly ovšem velmi zrychluje (a hodiny nudné práce krátí) procedura *Codebook*.

Po kliknutí na tlačítko *Continue* (viz obr. 3.1b) a pak na *OK* (viz obr. 3.1a) získáme následující výstupy (viz výstupy 3.1a až 3.1c):

Výstupy z operace *Frequencies*

Statistics

		pohlavi	roknar Rok narození
N	Valid	180	180
	Missing	0	0
Minimum		0	1879
Maximum		3	1991

Výstup 3.1a

pohlavi

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 muz	95	52,8	52,8	52,8
	1 zena	79	43,9	43,9	96,7
	2	3	1,7	1,7	98,3
	3	3	1,7	1,7	100,0
	Total	180	100,0	100,0	

Výstup 3.1b

roknar Rok narození

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1879	1	,6	,6	,6
	1947	1	,6	,6	1,1
	1973	1	,6	,6	1,7
	1974	3	1,7	1,7	3,3
	1975	2	1,1	1,1	4,4
	1976	9	5,0	5,0	9,4
	1977	15	8,3	8,3	17,8
	1978	18	10,0	10,0	27,8
	1979	43	23,9	23,9	51,7
	1980	58	32,2	32,2	83,9
	1981	27	15,0	15,0	98,9
	1990	1	,6	,6	99,4
	1991	1	,6	,6	100,0
	Total	180	100,0	100,0	

Výstup 3.1c[64]

[64] Všimněte si, že tyto výpočetní produkty SPSS označujeme výrazem „výstupy", a ne „tabulky", byť v některých případech tabulku připomínají. Děláme to z toho důvodu, že to, jak má vypadat tabulka, je v českém prostředí zcela jasně definováno a výstup z SPSS této definici neodpovídá. Jak má správná tabulka vypadat, si ukážeme později (viz kapitolu 8).

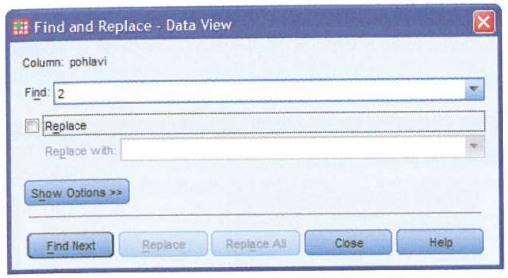
Při čištění dat začneme tím, že zkontrolujeme **počet jednotek**. K tomu slouží výstup 3.1a. Vidíme v něm především, že u obou proměnných je počet případů s uvedenou odpovědí 180 (v prvním řádku nazvaném *N Valid*). To je v pořádku, neboť přijímacího řízení se skutečně zúčastnilo 180 uchazečů. Kontrola celkového počtu je vždycky velmi důležitá. Pokud bychom našli příliš mnoho chybějících údajů (*missing values* – viz druhý řádek), znamenalo by to samo o sobě důležitou informaci, že něco není s příslušnými proměnnými v pořádku a je třeba zjistit, proč tam chybějící hodnoty jsou.

Dále se díváme na **rozsah hodnot**. Vidíme (stále ještě ve výstupu 3.1a), že u proměnné „pohlaví" je minimální hodnota 0 a maximální 3, což je zřetelně omyl, neboť interval, v němž se hodnoty této proměnné mohou pohybovat, je <0; 1>. Podobně u proměnné „rok narození" je minimální a maximální hodnota mimo reálný rámec. Ve výstupu 3.1b máme rozložení proměnné pohlaví. Vidíme, že pohlaví s hodnotou „2" mají tři a hodnotu „3" mají rovněž tři případy. Šest případů bylo tedy kódováno chybně, a je proto nutné vyhledat původní formuláře (přihlášky ke studiu), z nichž byla data pořízena, a chyby opravit. Ve výstupu 3.1c je rozložení hodnot roku narození. Jeden uchazeč má rok narození 1879 (jelikož se jedná o údaje z roku 1998, je to 119letý uchazeč o studium, což je očividně chybný údaj) a jeden 1991 (tedy 7letý uchazeč – chybný údaj, nebo génius?). Tyto údaje je opět nutné po kontrole s údaji ve formuláři opravit. Jeden uchazeč se narodil v roce 1947 – i to je asi omyl, neboť se jedná o jedenapadesátiletého uchazeče o prezenční (denní) studium. Ale zcela jisti si v tomto případě být nemůžeme (proč by se člověk zralého středního věku nemohl hlásit k dennímu studiu?), proto i tento údaj zkontrolujeme.

Nyní tedy víme, že v našem datovém souboru jsou chyby, které je třeba opravit. Máme dvě možnosti. Pokud máme dostatečně velký soubor (např. kolem 2 000 respondentů), můžeme si klidně dovolit těchto několik chybných případů obětovat a chybné hodnoty prohlásit (to je rekódovat) za hodnoty chybějící (*missing values*) – jak to udělat, si ukážeme v kapitole 6. *Missing values* pak nevstupují do žádných analýz. Máme-li relativně malý soubor (do tří čtyř stovek), měli bychom chyby opravit podle skutečných hodnot.[65]

Vyhledat chybu není příliš obtížné. Hledáme ji přímo v datech, v datovém editoru (*Data View*). Postupujeme následovně: V datovém editoru klikneme na proměnnou, v níž hledáme chyby. V našem případě to je proměnná *pohlavi*. Klikneme tedy na ni, aby se celý sloupec barevně zvýraznil. Pak klikneme na tlačítko *Edit* a ve vyskočivším okně klikneme na *Find* (nebo stiskneme klávesy Ctrl+F). Do příslušného okénka vepíšeme chybnou hodnotu, kterou chceme nalézt. My budeme nejdříve hledat hodnotu „2" (viz obr. 3.2).

[65] Doporučujeme ovšem, abychom chyby opravovali i ve velkých souborech. Sběr dat je finančně velmi nákladný a každý údaj, který je nevyužit, je velkým plýtváním penězi daňových poplatníků (neboť peníze na výzkum, které získáváme z vědeckých grantů, jdou koneckonců z daní nás všech).



Obr. 3.2 Hledání chybné hodnoty

Klikneme myší na *Find Next* – ve zvýrazněném datovém sloupci se objeví zvýrazněná buňka s hodnotou „2". Podíváme se do sloupce ID (identifikace, většinou je to první sloupec naší datové matice), abychom zjistili, o který případ se jedná. V našem souboru je to uchazeč č. 17. Není nyní nic lehčího, než jít do formuláře z přijímacího řízení, vyhledat uchazeče s číslem 17 (jistě je máme všechny dobře archivovány a seřazeny podle čísla identifikace), zjistit, jakého je pohlaví, a do vysvíceného políčka vepsat správnou hodnotu. A pokračujeme dále – jelikož víme, že v datech byly hodnoty „2" celkem třikrát, klikneme opět na *Find Next*, zjistíme identifikační číslo uchazeče a hodnotu „2" opravíme. Totéž pak uděláme ještě potřetí. Stejným způsobem opravíme v proměnné pohlaví i chybné hodnoty „3".

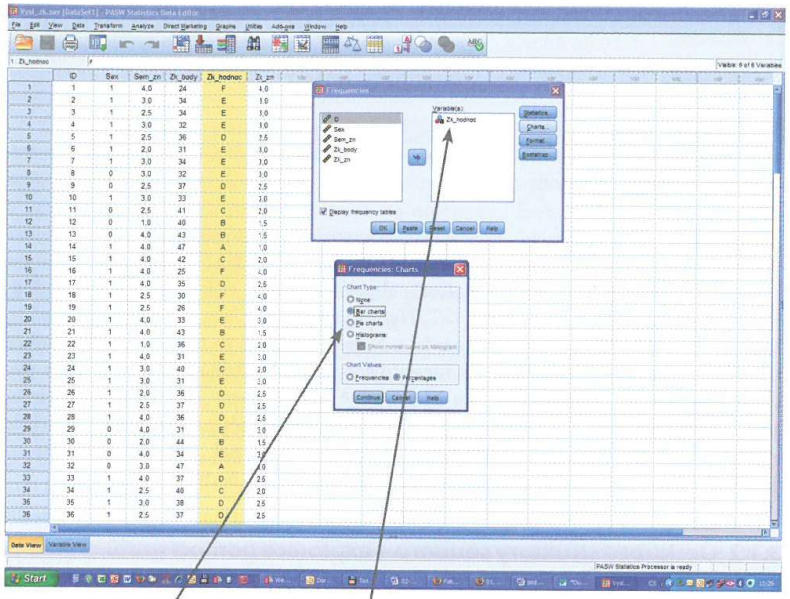
3.1.2 Deskripce struktury souboru – explorace pomocí grafů

Až poté, kdy jsme zkontrolovali všechny proměnné v souboru a data vyčistili, můžeme přistoupit k vlastní analýze. Jak jsme již zdůraznili v úvodu této kapitoly, začínáme vždy jednorozměrnou analýzou, tříděním podle jedné proměnné, tříděním prvního stupně. To nám dá představu o rozložení jednotlivých proměnných a umožní nám získat základní představu o tendencích v datech. Pro ještě lepší představu doporučujeme nechat si udělat i příslušné grafy. Upozorňujeme ovšem dopředu, že vytváření grafů v SPSS je vhodné pro naše vlastní porozumění. Pokud bychom však chtěli rozložení proměnné prezentovat v nějakém výstupu, např. v prezentaci na přednášce nebo v publikovaném textu, nejsou grafy z SPSS nejvhodnější, takže doporučujeme data vkopírovat do Excelu nebo Power Pointu a graf vytvořit s jejich pomocí. Pro usnadnění práce připomeňme, že výstup z *Frequencies* lze pomocí Ctrl+C a Ctrl+V zkopírovat a vložit přímo do Excelu, takže není nutné hodnoty přepisovat.

Třídění prvního stupně a zobrazení distribuce hodnot jedné proměnné získáme opět prostřednictvím procedury *Frequencies* a její grafické zobrazení kliknutím na tlačítko *Charts*.

Příklad 3.1

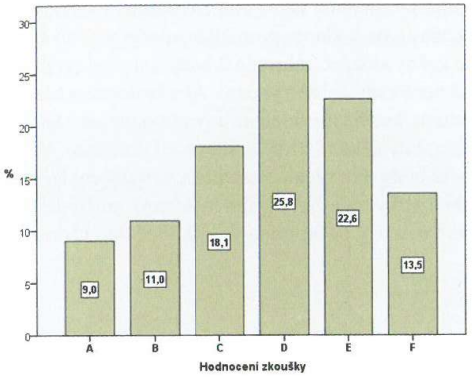
Chceme znát, jaké byly výsledky z předmětu metody výzkumu v sociologii (proměnná *Zk_hodnoc* v souboru „Vysl-zkousky").



Obr. 3.3 Vytvoření sloupcového grafu kategorizované proměnné

Řešení: Do dialogového okna *Frequencies* vložíme proměnnou *Zk_hodnoc* a po kliknutí na tlačítko *Charts* máme možnost zvolit tři druhy grafů: *Bar*, tedy **sloupcový graf**, *Pie*, tedy **graf koláčový** a *Histogram*, tedy **histogram četností** (viz obr. 3.3). Pro kategorizované proměnné, to je pro proměnné nominální a ordinální, jsou určeny grafy sloupcové nebo koláčové (pozor ale, koláčového grafu lze použít pouze tehdy, když zobrazujeme všechny kategorie, které tvoří dohromady jeden logický celek), pro kardinální proměnné je určen histogram. Naše proměnná *Zk_hodnoc* je proměnnou ordinální (byť textovou, nečíselnou, s alfabetickými kódy A–F, což je hodnocení, které používá Masarykova univerzita, kdy výsledek A je nejlepší, F znamená neúspěch), proto v příslušném dialogovém okně necháme udělat sloupcový graf (to je zaškrtneme políčko *Bar charts*). Rozhodneme se také, zdali chceme, aby v grafu byly na ose Y absolutní četnosti (počty studentů), nebo procenta (relativní četnosti). Doporučujeme pracovat vždy raději s procenty. Vše je znázorněno na obr. 3.3. Výsledek pak v grafu 3.1 (všechny grafy jsou na rozdíl od výstupů, které se vám zobrazí, pro lepší přehlednost lehce editovány).

Graf 3.1 jasně ukazuje, že většina studentů byla ve zkoušce hodnocena známkou D nebo E a že nejlepší hodnocení A získalo méně studentů, než bylo těch, kdo u zkoušky neuspěli.[66]



Graf 3.1 Výsledky zkoušky z předmětu metody výzkumu v sociologii[67]

Pozn. Grafy lze editovat tak, že dvakrát klikneme na obrázek. Tím se dostaneme do editovacího režimu, jenž umožňuje graf nejrůznějším způsobem upravovat. Není to však příliš snadné, a jak jsme již řekli, pro vytváření grafů doporučujeme raději Excel.

Chceme-li graficky prozkoumat (cxplorovat) kardinální proměnnou, musíme jako typ grafu zvolit histogram četností. Postup je víceméně stejný jako v případě proměnných kategorizovaných (viz obr. 3.3), jen se dvěma rozdíly. Především, pochopitelně, namísto sloupcového grafu (*Bar charts*) zaškrtneme kliknutím políčko *Histograms*. A dále, jelikož kardinální proměnné mají obvykle dlouhou řadu hodnot, budeme požadovat, aby se při vytváření grafu netvořil výstup rozložení četností. Dosáhneme toho tak, že v dialogovém okně *Frequencies* (viz obr. 3.3) zrušíme zaškrtnutí okénka *Display frequencies table* (zobraz tabulku četností).

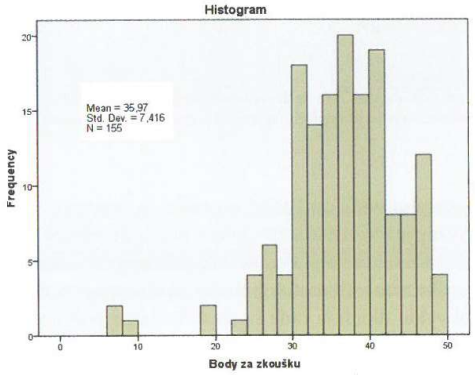
[66] Stejný graf je možné získat i prostřednictvím procedury *Graphs* (*Graphs – Legacy Dialogs – Bar – Simple – Summaries of groups of cases – Define*). Zájemce o to, jak v SPSS Statistics pracovat s grafy, odkazujeme na 4. kapitolu učebnice Field, A. (2009). *Discovering Statistics Using SPSS* (3rd ed.). London: Sage, v níž je vše popsáno do detailu. V naší učebnici se grafům věnujeme pouze informativně.

[67] U této proměnné by asi bylo zajímavé zobrazit výsledky uspořádané podle četností – někdy je prostě z nejrůznějších důvodů užitečné vytvořit graf takovým způsobem, aby jeho jednotlivé sloupce byly uspořádány sestupně od nejvyššího po nejnižší. V Excelu ho vytvoříme tak, že hodnoty zobrazované proměnné necháme v datovém listu setřídit sestupně podle velikosti a z takto setříděných hodnot si necháme udělat sloupcový graf.

Příklad 3.2

Nechejme si graficky zobrazit rozložení proměnné bodového výsledku u zkoušky (*Zk_body*).

Řešení: Výsledek ukazuje graf 3.2. Abyste si uvědomili, jak se liší histogram od sloupcového grafu, udělejte si pro tuto proměnnou také sloupcový graf. Pak zjistíte, že histogram nezobrazuje každou jednotlivou hodnotu proměnné, ale že vytváří třídy hodnot (intervaly), v našem případě každý sloupec zahrnuje 2 body, interval program „vymyslel" automaticky za nás (což není vždy úplně výhoda). Aby histogram naznačoval, že se jedná o spojitou proměnnou, nedělá mezi jednotlivými hodnotami na ose X žádnou mezeru, s výjimkou případu, kdy některá třída hodnot není obsazena. V našem příkladu to je např. interval 20–22 body. Pro rychlou orientaci v rozložení hodnot program tiskne také základní charakteristiky rozložení: průměr (*Mean*), směrodatnou odchylku (*Standard Deviation*) a také celkový počet případů (N). O těchto charakteristikách se zmíníme za chvíli.

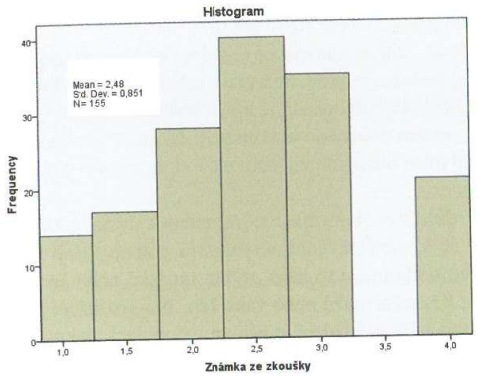


Graf 3.2 Histogram bodového hodnocení zkoušky

Když se díváme na histogram, zajímáme se o celkový tvar rozložení, to je zdali má jeden nebo více vrcholů a zdali má víceméně symetrický tvar, nebo je naopak na nějakou stranu vychýlený – vychýlený může být doleva nebo doprava. Dále se zajímáme o výrazné odchylky od tohoto rozložení, především o odlehlé hodnoty – jelikož máme již jako správní výzkumníci data vyčištěna, nejsou případné odlehlé hodnoty výsledkem chyby, ale reálného chování proměnné (lépe řečeno chování nositelů této vlastnosti). V našem případě vidíme, že rozložení je výrazně nepravidelné, neboť má více vrcholů. Nemá tedy symetrický tvar a ani nelze říci, na kterou stranu je vychýlené. Vidíme ale, že má dvě výrazně odlehlé hodnoty, několik studentů získalo méně než 10 bodů z celkových 50 možných.

Tvar rozložení se změní, když bodový zisk u zkoušky přetavíme na novou proměnnou, a to „známku u zkoušky", která bude nabývat standardních stupňů hodnocení, které se

používají na Masarykově univerzitě: 1; 1– (1,5); 2; 2– (2,5); 3 a 4 (= neprospěl/a). Provedeme to tak, že bodový zisk transformujeme do intervalů, jimž přiřadíme jednotlivé hodnotící stupně: 0–30 bodů = 4; 31–34 b. = 3; 35–38 b. = 2,5; 39–42 b. = 2; 43–46 b. = 1,5; 47–50 b. = 1. Tento postup transformace proměnné (tzv. rekódováním) si ukážeme později. V datovém souboru „Vysl_zk.sav" je tato nová proměnná již vytvořena, takže ji můžeme graficky zobrazit (viz graf 3.3).



Graf 3.3 Histogram četností známky u zkoušky

Pozn. Histogram jsme vytvořili pouze z didaktických důvodů, v reálném analytickém životě bychom namísto histogramu zvolili sloupcový graf, neboť provedenou transformací jsme původní kardinální proměnnou, která měla velký počet kategorií, změnili na proměnnou s malým počtem kategorií, takže vznikla de facto proměnná kategorizovaná. A u ní histogram nemá valný smysl.

Ve srovnání s grafem 3.2 se rozložení v gafu 3.3 dosti odlišuje. Především má opačně orientovanou osu X. Zatímco v grafu 3.2 byli studenti s nízkým bodovým ziskem v levé části osy X, zde jsou v její pravé části, neboť nízký bodový zisk znamená vyšší (a tedy horší) známku. Rozložení již nemá nepravidelný tvar s mnoha vrcholy, naopak má jen jeden vrchol a blíží se symetrickému rozložení s mírným (ale opravdu jen mírným) vychýlením doleva (prázdný sloupec u hodnoty 3,5 je způsoben tím, že takovou známku systém hodnocení na MU nezná, takže je možné si představit, že sloupec známky 4 by mohl být na této pozici). Jelikož jsme v tomto grafu vytvořili širší intervaly pro zobrazení, získali jsme sice elegantnější tvar, ale ztratili jsme část důležitých informací (hovoříme o redukci informace). Vypadl například údaj o tom, že v souboru máme několik studentů, kteří se na zkoušku nedokázali vůbec připravit (anebo je pedagogické schopnosti jejich učitele nedokázaly pro tento předmět nadchnout – tuto kritiku si můžeme dovolit, neboť oněmi učiteli jsou autoři této učebnice).

Pro jakoukoliv explorační analýzu prostřednictvím grafů platí, že volba měřítka grafu nebo zobrazovaného intervalu vždy ovlivňuje výslednou podobu grafu, takže

jeho „okometrický“ rozbor může vést k nepřesným závěrům. Na to si musíme dávat vždy velký pozor nejen u své vlastní analýzy, ale především u kontroly výsledků práce jiných. Pamatujme na to, že **grafy mohou být nejen ilustrativnější než čísla, ale někdy také zavádějící a zkreslující** (blíže k tomu například kapitola „Správné a nesprávné používání grafů“ v knize *Statistika pro obchod a hospodářství*).[68] Volbou jejich podoby lze poměrně snadno docílit toho, aby představovaly sdělení, které vyhovuje tvůrci grafu a jeho záměrům. Této skutečnosti jsou si velmi dobře vědomi např. politici (pouhou změnou měřítka grafu lze například změnit pozvolný trend představující praktické zachycení stagnace v obraz růstu apod.).

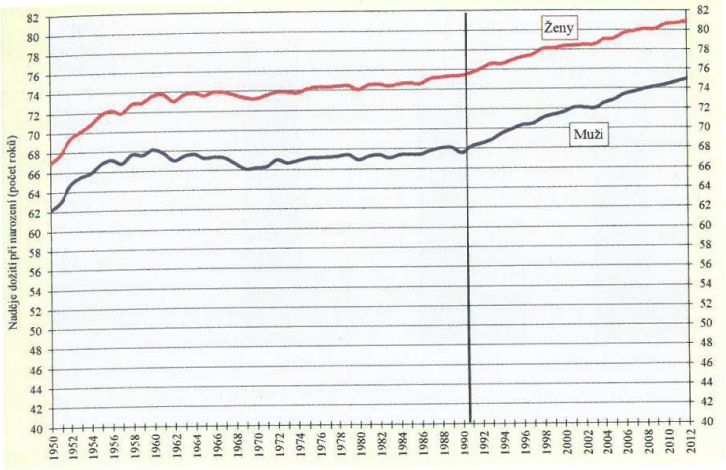
Vraťme se nazpět k našemu příkladu. Jelikož numerické (číselné) hodnocení výsledku zkoušky odpovídá jeho alfabetické podobě, je jasné, že tvar rozložení v grafu 3.3 musí být totožný s tvarem v grafu 3.1. Drobný vizuální rozdíl je způsoben tím, že graf 3.1 je grafem sloupcovým a že zobrazuje procenta namísto absolutních četností. Srovnáním těchto dvou grafů se také ukáže rozdíl mezi histogramem četností a sloupcovým grafem.

Celkový tvar rozložení skýtá důležitou informaci o proměnné. Mnohé lidské vlastnosti, tedy proměnné, s nimiž se v sociálněvědním výzkumu pracuje, mají tvar rozložený symetricky kolem jednoho vrcholu. Je to např. výška mužské nebo ženské populace, inteligence měřená testem IQ, váha mužů nebo váha žen. Jiná rozložení jsou asymetrická, vychýlená doprava – typickým příkladem je příjem v populaci. Pokud má rozložení více než jeden vrchol, může to indikovat heterogenitu souboru (váha české populace by měla jistě dva vrcholy, jeden vrchol pro muže, druhý pro ženy). Na začátku každé analýzy stojí rozhodně za to u každé proměnné studovat rozložení jejích hodnot.

Důležitým typem explorační analýzy jsou **analýzy trendů**. Ty předpokládají, že máme údaje o chování nějaké proměnné v delší časové řadě. Sociologický výzkum data takového druhu obvykle neobsahuje, neboť je většinou jednorázovým šetřením v jednom časovém okamžiku (tzv. transverzální výzkum). Opakované výzkumy, v nichž se kladou tytéž otázky[69], nejsou jako jedna z forem longitudinálních výzkumů bohužel v sociálních vědách příliš časté. Slovo „bohužel“ je zde zcela namístě, neboť teprve časová řada umožňuje lépe pochopit chování dané proměnné a naznačit příčiny její variability. Opakované výzkumy jsou ovšem finančně i organizačně náročné, proto proměnné v časových řadách získáváme spíše z dat statistických úřadů (existence takových časových řad ovšem nezávisí na záměrech výzkumníků, ale na potřebách státní správy). Přesto i zde můžeme najít velmi zajímavá data. Časová řada založená na datech Českého statistického úřadu, zachycující vývoj naděje dožití české populace (viz graf 3.4), je toho dokladem.

[68] Wonnacot, T. H., & Wonnacot, R. J. (1993). *Statistika pro obchod a hospodářství*. Praha: Victoria Publishing.

[69] Klasickými reprezentanty tohoto typu výzkumů jsou například European Social Survey nebo European Values Study (jehož data používáme v této publikaci).

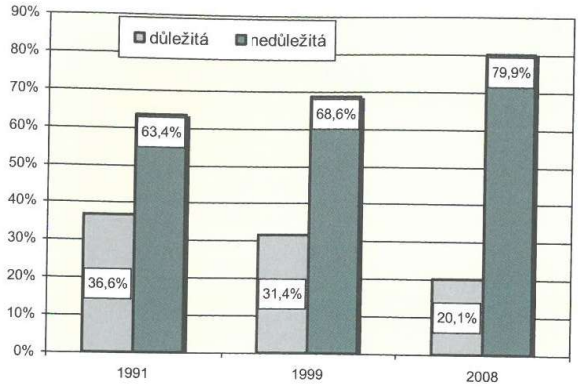


Graf 3.4 Naděje dožití při narození českých mužů a žen 1950–2012

Graf 3.4 pěkně dokládá, jak se změna politického, ekonomického a sociálního režimu po roce 1990 projevila příznivě na vzorci české úmrtnosti. U mužů i u žen se od té doby každoročně začala téměř lineárně zvyšovat pravděpodobnost, že se v průměru dožijí delšího věku.

V některých případech ale máme k dispozici i časovou řadu ze sociologických dat. V Evropě probíhá od roku 1981 výzkum, který má zachytit proměnu hodnot a postojů k nejrůznějším aspektům a fenoménům života společnosti. Nazývá se European Values Study (EVS). Dosud proběhly jeho čtyři vlny, a to v letech 1981, 1990–91, 1999, 2008. Česká republika se k tomuto nejrozsáhlejšímu evropskému srovnávacímu sociologickému výzkumu přidala v roce 1991 (dříve to z politických důvodů nebylo možné). Jelikož se velká část otázek v tomto výzkumu opakuje, máme tak pro řadu proměnných časové řady. Ukázku jedné z nich uvádí graf 3.5, na němž je patrný vývoj důležitosti politiky v každodenním životě.[70]

[70] Graf byl vytvořen na základě výpočtů SPSS prostřednictvím Excelu.



Graf 3.5 Důležitost politiky v každodenním životě

Zdroj: EVS ČR 1991–1999–2008.

Data grafu jasně říkají, že politika ztratila od r. 1991 pro českou populaci důležitost: z 37 % respondentů, kteří v roce 1991 uvedli, že politika je pro ně v životě důležitá, klesl do r. 2008 tento podíl na 20 %. Zrcadlově s tím se pochopitelně zvýšily podíly respondentů, pro něž je politika nedůležitá. Jistě stojí za další analýzy, proč k tomuto vývoji dochází, když s modernizací společnosti by měl jít ruku v ruce také rozvoj občanské společnosti, v níž jsou politické postoje a aktivity na lokální úrovni její organickou součástí.

Grafická analýza ovšem pro úplné pochopení proměnlivosti (variability) hodnot proměnné nepostačuje. Dává sice první a důležité indikace, avšak ne všechny. Z tohoto důvodu doplňujeme jednorozměrnou analýzu ještě analýzou prostřednictvím čísel a prostřednictvím výpočtů souhrnných charakteristik, jimiž jsou především tzv. **míry polohy** a **míry variability (rozptýlenosti)**.

3.2 Popis rozložení proměnných prostřednictvím čísel

Statistické rozložení **kategorizovaných** proměnných může být vyjádřeno v tabulce, která obsahuje absolutní četnosti, relativní četnosti a kumulativní relativní četnosti. Pro proměnné **kardinální** (spojité) si obvykle žádnou tabulku dělat nenecháváme, tyto číselné charakteristiky nesledujeme a pracujeme pouze s jejich souhrnnými charakteristikami. Je to z toho důvodu, že to jsou obvykle proměnné s velkým počtem hodnot (např. věk), takže pokud bychom chtěli takovou proměnnou uspořádat do tabulky a v našem souboru měli např. populaci ve věku 20–65 let, měla by tabulka 46 řádků.

Absolutní četnosti (v jazyce SPSS *Frequencies*) udávají, kolik případů mají jednotlivé kategorie proměnné. Například kolik je v souboru mužů (žen) nebo kolik je v souboru osob s jednotlivými stupni dosaženého vzdělání. Součet absolutních

četností ve všech kategoriích (včetně chybějících hodnot) dává celkové *n*, to je rozsah (velikost) souboru.

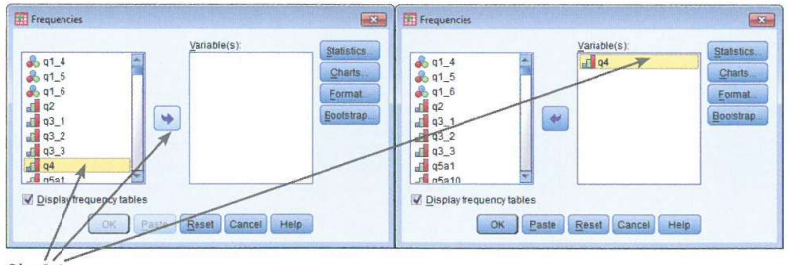
Relativní četnosti (*Percent*) ukazují, jaký podíl představují v celém souboru případy mající danou vlastnost z celkového počtu jednotek v souboru. Obvykle se násobí stem, takže hovoříme o procentech. Například sledujeme, jaký je v celém souboru podíl mužů a žen nebo jaký je v souboru procentuální podíl osob s jednotlivými stupni dosaženého vzdělání. V případě, že zjišťujeme procentuální podíl pouze z celku případů s danou vlastností, tedy když nejsou do tohoto celku zahrnuty případy, u nichž nemáme údaj, zdali vlastnost mají, nebo ne (jsou to případy s chybějící hodnotou neboli s *missing values*), získáváme **platná procenta** (*Valid Percent*). Součet relativních četností ve všech kategoriích dává 100 %.

Kumulativní relativní četnosti (*Cumulative Percent*) říkají, jaký podíl představují v souboru případy mající vlastnosti s nižší či stejnou hodnotou (jaký podíl například představují v souboru jedinci s maximálně středním vzděláním, což je kumulace jedinců se základním a středním vzděláním apod.). Tento druh procent se používá jen v případě ordinálních nebo intervalových proměnných – nemají smysl u proměnných nominálních, neboť zde je pořadí hodnot zcela arbitrární a jednotlivé kategorie nemá smysl kumulovat.

Příklad 3.3

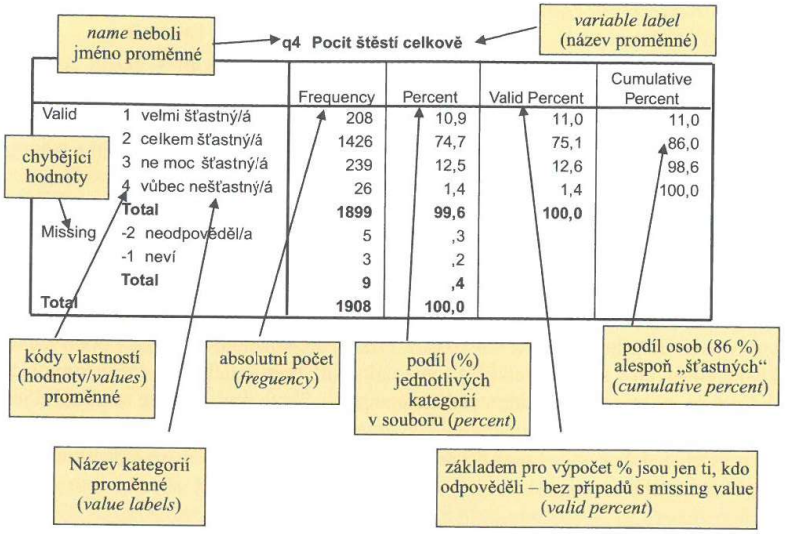
Ve výzkumu EVS (soubor „EVS99-cvicny") nás zajímá rozložení proměnné pocit štěstí (*q4*).

Řešení: V menu *Analyze – Descriptive Statistics – Frequencies* proměnnou, jejíž rozložení chceme zobrazit, vybereme vysvícením a kliknutím na tlačítko šipky mezi okny ji přesuneme napravo. Lze tak zadat i více proměnných a jen praktické důvody obvykle brání tomu, abychom zadali všechny proměnné najednou. Vše ilustruje obrázek 3.4.



Obr. 3.4

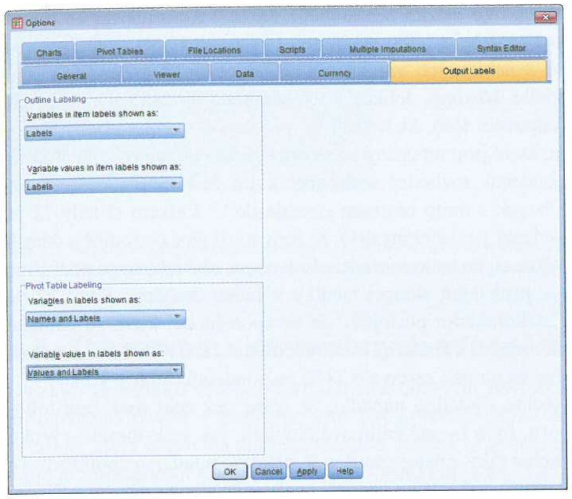
Základním výstupem procedury *Frequencies* je frekvenční tabulka (viz výstup 3.2). Jak této tabulce rozumíme? Popisky na obr. 3.5 vše vysvětlují. Slovní výklad toho, jak lze čísla z třídění I. stupně „číst", si ukážeme na dalším příkladu.



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 velmi šťastný/á	208	10,9	11,0	11,0
	2 celkem šťastný/á	1426	74,7	75,1	86,0
	3 ne moc šťastný/á	239	12,5	12,6	98,6
	4 vůbec nešťastný/á	26	1,4	1,4	100,0
	Total	**1899**	**99,6**	**100,0**	
Missing	-2 neodpověděl/a	5	,3		
	-1 neví	3	,2		
	Total	**9**	**,4**		
Total		**1908**	**100,0**		

Výstup 3.2 Rozložení četností proměnné „pocit štěstí"

Poznámka: Abychom dostali ve výstupech z početních operací tabulky, v nichž budou jak kódy (nebo hodnoty) proměnné (*values*), tak i popisky těchto kódů (*labels*), tedy ve formě „1 velmi šťastný", musíme si v *Edit – Options – Output Labels* nastavit prostředí tak, jak vidíte na obrázku 3.5, to je nastavit v *Pivot table labelling* formát *Names and Labels* a *Values and Labels*.



Obr. 3.5 Nastavení vnitřního prostředí SPSS pro zobrazení jména proměnné a jejího popisu

Příklad 3.4

V mezinárodním komparativním výzkumu European Values Study, který v České republice provedl v roce 1999 Jan Řehák a Ladislav Rabušic (data sbírala agentura SC&C) na reprezentativním souboru české dospělé populace (ve věku 18 let a starším) byla mimo jiné také položena otázka: „Lidé hovoří o měnících se rolích dnešních mužů a žen. Řekněte nám nyní, nakolik souhlasíte s následujícím výrokem: *Zaměstnání je dobrá věc, po čem však většina žen opravdu touží, je domov a děti.*" Respondenti měli možnost s výrokem „rozhodně souhlasit", „souhlasit", „nesouhlasit" nebo „rozhodně nesouhlasit". Zajímá nás nyní, jak byly odpovědi na tuto otázku v jednotlivých variantách rozloženy (datový soubor „EVS99-cvicny"). Byla to otázka, která měla v datové matici kód *q46_3*. Získali jsme tento výstup (3.3):

Q46_3 Většina žen touží po domově a dětech

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 rozhodně souhlasí	213	11,2	12,0	12,0
	2 souhlasí	1070	56,1	60,1	72,1
	3 nesouhlasí	474	24,9	26,6	98,7
	4 rozhodně nesouhlasí	22	1,2	1,3	100,0
	Total	1780	93,3	100,0	
Missing	-2 neodpověděl/a	8	,4		
	-1 neví	120	6,3		
	Total	128	6,7		
Total		1908	100,0		

Výstup 3.3 Rozložení četností proměnné sledující souhlas nebo nesouhlas s výrokem *Zaměstnání je dobrá věc, po čem však většina žen opravdu touží, je domov a děti.*

Tabulka SPSS říká, že na tuto otázku z celkového počtu 1 908 dotázaných odpovědělo 93,3 % respondentů (což bylo 1 780 osob) a 6,7 % respondentů (128 osob) se nerozhodlo ani pro jednu z nabízených variant (jejich odpovědi chybějí, proto jsou v tabulce umístěny do oddílu *Missing*). Jelikož z výzkumného hlediska mají pro nás většinou význam pouze odpovědi těch, kteří mají na položenou otázku nějaký názor, pracujeme obvykle s údaji, které jsou umístěny ve sloupci *Valid Percent* (platná procenta). Vidíme, že 12 % respondentů „rozhodně souhlasilo" s tím, že ženy především touží po domově a dětech. 60 % pak s tímto názorem „souhlasilo".[71] Celkem si tedy 72 % (12 + 60,1) českých respondentů v r. 1999 myslelo, že ženy touží více po rodině a dětech než po zaměstnání (všimněme si, že tento výsledek dostaneme také tak, že se podíváme na kumulativní procento v posledním sloupci tabulky v řádku druhém). Podíl 72 % osob s tímto genderově „nekorektním postojem" je ve společnosti, která se tradičně vyznačuje silnou zaměstnaností žen a poměrně i silnou feministickou rétorikou, poněkud překvapivé zjištění. Opačný názor pak zastávalo 28 % respondentů (26,6 + 1,3).

Pokud vás při čtení těchto výsledků napadlo, že spíše než znát rozložení tohoto postoje v celém souboru, bylo by asi zajímavější zjistit, jak se k tomuto výroku staví ženy a jak muži nebo jaký postoj zaujímají věkově mladší respondenti ve srovnání s těmi věkově staršími, pak jste na správné analytické stopě a tíhnete k třídění II. stupně. Skutečně, třídění I. stupně neboli četnostní tabulky nemají v sociologických analýzách příliš velký věcný význam. Jsou ovšem, jak jsme již ukázali, neocenitelným pomocníkem při kontrole dat a také dobře slouží jako základní informace před složitějšími analýzami. A samozřejmě, jsou např. hlavním typem výstupů ve výzkumech veřejného mínění, které nám třeba sdělují: „Pokud by se volby konaly příští týden, k voličským urnám by se dostavilo 62 % voličů. 12 % voličů je přesvědčeno, že volit nepůjde, zbylých 26 % je zatím nerozhodnutých."

Na závěr této pasáže si dovolujeme jedno důležité pedagogické upozornění.

Tabulkové výstupy z SPSS, jakkoliv jsou vhodné pro analytické účely, není možné nikdy bez úprav použít v textech k publikaci (v článku, diplomové práci) nebo k prezentaci (na semináři či konferenci). Nesplňují totiž požadavky na to, jak má správná tabulka vypadat.

Vzor takové tabulky, jež je upraveným výstupem 3.4, je uveden níže (viz tab. 3.1). Každá tabulka musí mít své číslo, v textu je číslujeme průběžně. Musí mít svůj název, musí mít popsány jednotlivé kategorie odpovědí (jednotlivé varianty proměnné), čísla uvádíme většinou v procentech a v posledním řádku uvedeme i údaj o celkovém počtu jednotek, aby si případný zájemce mohl jednoduchým výpočtem, kdyby to potřeboval, převést uváděná procenta na absolutní četnosti. Každá tabulka musí mít uveden zdroj, z něhož data pocházejí.

[71] Přesněji řečeno, bylo jich 60,1 %, ale procenta vždy zaokrouhlujeme na celá čísla – uvádění procent na desetinná místa totiž předstírá přesnost, která v datech pocházejících ze *survey* zdaleka není.

	%
Rozhodně souhlasí	40
Souhlasí	22
Nesouhlasí	8
Rozhodně nesouhlasí	11
n = 1 837	100

Zdroj: EVS ČR 1999.

Tab. 3.1 Rozložení četností proměnné sledující souhlas nebo nesouhlas s výrokem *Zaměstnání je dobrá věc, po čem však většina žen opravdu touží, je domov a děti*, ČR 1999

3.3 Zpracování vícenásobných odpovědí

Otázky v dotazníku většinou formulujeme takovým způsobem, že respondenta žádáme o jednu odpověď. Chceme po něm, aby z předem definovaných kategorií odpovědí zvolil tu, která odpovídá jeho skutečnosti. Někdy se ovšem musíme uchýlit k otázce, v níž je možné volit i odpovědí více. Výsledkem je tzv. vícenásobná odpověď (*multiple response* či MR), která je v datech zaznamenána jako několik proměnných. Například v jednom reprezentativním výzkumu, který se mimo jiné týkal i problematiky populační politiky a který jsme na populaci ve věku 20–45 let provedli v roce 2005, jsme položili následující otázku (pracujeme se souborem „M-P-R.sav")[72]:

F8: Můžete nám, prosím, říci, jaké jsou důvody toho, že si nepřejete (další) dítě? Z následujícího seznamu vyberte tři důvody, které jsou pro Vás nejvíce důležité, a tyto důvody uspořádejte podle pořadí důležitosti.

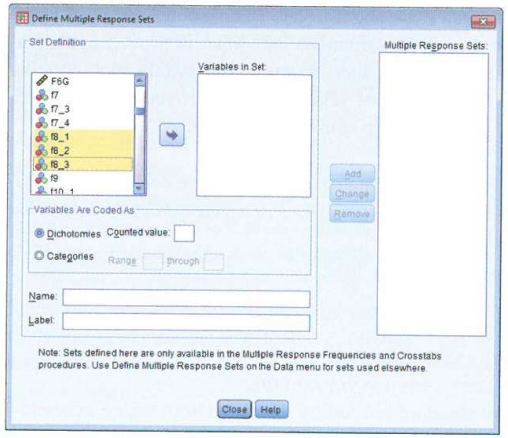
1. Už mám tolik dětí, kolik jsem chtěla.
2. Nedovoluje to můj zdravotní stav.
3. Žiji sama a nemám stálého partnera.
4. Moje práce a profesní aktivity to neumožňují.
5. Musela bych obětovat čas, který věnuji svým zájmům.
6. Ohrozilo by to vážně životní standard mé rodiny / měli bychom vážné existenční potíže.
7. Mám-li být upřímná, myslím si, že bych nebyla dobrou matkou.
8. Příliš se obávám toho, jaká budoucnost by moje děti čekala.
9. Znamenalo by to pro mne riziko nezaměstnanosti / ztráty zaměstnání / zhoršení pracovní pozice.
10. Nemohla bych si užívat života tak jako dosud.

[72] Výzkum byl financován Grantovou agenturou České republiky pod *názvem Rodina, práce a reprodukční strategie aneb preferenční teorie v ČR*. Řešiteli projektu byli Ladislav Rabušic a Beatrice Chromková Manea. M-P-R je akronym pro sousloví Manželství–Práce–Rodina. Dotazník z tohoto výzkumu je přiložen na CD.

11. Jsem na děti už příliš stará.
12. Můj partner je na děti už příliš starý.
13. Mít děti dnes nemá žádný význam.
14. Můj partner nechce mít (další) dítě.
15. Porod a rodičovství jsou náročné.

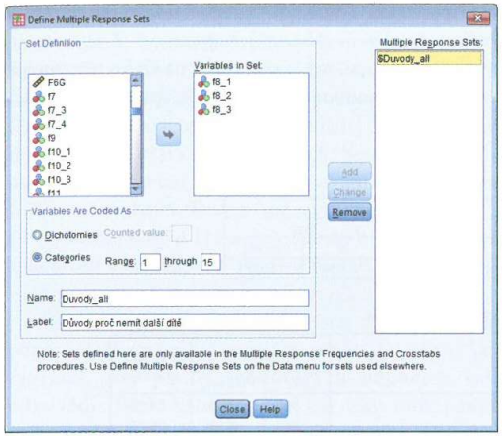
Tato proměnná byla do matice nahrána v podobě tří proměnných: jako *f8_1*, která zaznamenávala výběr na prvním místě, a *f8_2* a *f8_3* jako výběry na místě druhém a třetím. Analýza prostřednictvím třídění I. stupně nemá v tomto případě smysl, neboť bychom museli analyzovat každou proměnnou zvlášť a výsledný obraz tří voleb bychom dostali jen obtížně.[73] SPSS má pro tuto situaci speciální proceduru *Analyze – Multiple Response*. Funguje následovně.

Po kliknutí na *Multiple Response* ještě klikneme na *Define Variable Sets*, kde se otevře dialogové okno (viz obr. 3.6a). V něm musíme vyplnit všechna prázdná okna takovým způsobem, jak ukazuje obr. 3.6b.



Obr. 3.6a Výběr proměnných do zpracování mnohonásobných odpovědí

[73] Výjimkou je analýza první uvedené odpovědi (jakožto té nejdůležitější), neboť ta se často zpracovává. Anglická terminologie má pro tento fenomén označení *top of mind*, česky se zpravidla doslovně překládá jako „první na mysli".



Obr. 3.6b Zadání proměnných do zpracování mnohonásobných odpovědí

Především musíme proměnné, jejichž odpovědi budeme načítat a z nichž tedy vytvoříme novou proměnnou, vložit do okénka *Variables in Set*. V našem případě to jsou proměnné *f8_1*, *f8_2*, *f8_3*. Pak musíme SPSS říci, jakého typu jsou tyto proměnné, zdali dichotomie (*dichotomies*), nebo vícehodnotové proměnné (*categories*). My máme proměnné vícehodnotové, proto klinutím myši zaškrtneme *Categories* (viz obr. 3.6b). Současně musíme sdělit, jaký je obor hodnot našich proměnných. My jsme v dotazníku použili kódy 1–15, proto do okének *Range* a *through* vložíme číslici 1 a 15. Pak vyplníme technické jméno proměnné (*Name*), třeba jako *Duvody_all* a jeho popisek (*Label*). Nakonec klikneme na tlačítko *Add* (přidej) a nová proměnná se pod jménem *$Duvody_all* vepíše do okna *Multiple Response Sets*.[74]

[74] Dodejme, že v datech se žádná nová proměnná nevytvoří, takže jde de facto o virtuální konstrukt určený pro následnou analýzu.

Novou proměnnou teď máme připravenou a můžeme si nechat udělat její třídění.[75] Zobrazme si nyní četnostní tabulku: *Analyze – Multiple Response – Frequencies*. Dialogové okno nabídne naši novou proměnnou, my ji kliknutím na šipku přesuneme do okna nazvaného *Table(s) for* a dalším kliknutím na tlačítko *OK* spustíme třídění.[76] Zde je výsledek (viz výstup 3.4).

Case Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
$Duvody_all	**1465**	57,6%	1081	42,4%	2546	100,0%

$Duvody_all Frequencies

		Responses		Percent of Cases
		N	Percent	
$Duvody_all	1 Už mám tolik dětí, kolik jsem chtěl/a	800	18,7%	54,6%
	2 Nedovoluje to můj zdravotní stav	228	5,3%	15,5%
	3 Žiji sám/a a nemám stálého partnera/ku	325	7,6%	22,2%
	4 Moje práce a profesní aktivity to neumožňují	201	4,7%	13,7%
	5 Musel/a bych obětovat čas, který věnuji svým zájmům	248	5,8%	16,9%
	6 Měli bychom vážné existenční potíže	527	12,3%	36,0%
	7 myslím si, že bych nebyl/a dobrou matkou/otcem	127	3,0%	8,6%
	8 Příliš se obávám, jaká budoucnost by moje děti čekala	486	11,3%	33,2%
	9 Znamenalo by to pro mne riziko nezaměstnanosti/	210	4,9%	14,3%
	10 Nemohl/a bych si užívat života tak jako dosud	183	4,3%	12,5%
	11 Jsem na děti už příliš stará/ý	327	7,6%	22,3%
	12 Můj partner/ka je na děti už příliš starý/á	97	2,3%	6,6%
	13 Mít děti dnes nemá žádný význam	40	,9%	2,7%
	14 Můj partner/ka nechce mít (další) dítě	223	5,2%	15,2%
	15 Porod a rodičovství jsou náročné	263	6,1%	18,0%
Total		**4284**	100,0%	292,3%

Výstup 3.4 Třídění I. stupně proměnné založené na vícenásobné odpovědi na otázku „důvody proč nemít další dítě"

Výstup z výše uvedené četnostní tabulky je odlišný od standardních četnostních tabulek (od třídění I. stupně). Je to dáno tím, že údaje zde jsou dvojího druhu. Jednak údaje o respondentech, jednak údaje o počtu odpovědí. Jelikož každý respondent byl

[75] Ten, kdo si pozorně přečte poznámku (*Note*) na spodní straně dialogového okna, zjistí, že takto definovaná nová proměnná se dá použít pouze v proceduře *Multiple Response Set*, kde se nabízí pouze pro třídění I. (*Frequencies*) a II. stupně (*Crosstabs*). Pokud bychom tuto proměnnou chtěli používat i v jiném typu analýzy, museli bychom ji vytvořit jiným způsobem, a to v režimu *Data – Multiple Response Set*.

[76] Těm, kteří se chtějí naučit pracovat s SPSS prostřednictvím syntaktických příkazů, doporučujeme předtím kliknout na tlačítko *Paste* (vlož). Tím si jednak uchováte způsob, jak byla tato proměnná vytvořena, jednak máte připraveno zadání pro případný opakovaný výpočet. Soubor se syntaktickými příkazy je ovšem nutno uložit. Více o užívání příkazů (syntaxe) najde zájemce v dodatku II na konci knihy.

požádán, aby vybral tři důvody, měl by být počet odpovědí $n \times 3$. Vzhledem k tomu, že v souboru bylo celkem 2 546 respondentů, měli bychom získat 7 638 odpovědí. Jak ale říká první část výstupu 3.4 (nazvaná *Case Summary*), 1 081 respondentů na tuto otázku neodpovědělo. Je to pochopitelné, neboť této otázce předcházela otázka filtrační, zdali si respondent přeje další dítě. Pokud odpověděl kladně, to je že si další dítě přeje, na otázku po důvodech, proč *nechce* další dítě, jsme se ve výzkumu již samozřejmě neptali. Platných odpovědí (*Valid*) bylo celkem 1 465, což by mělo dávat $1\,465 \times 3 = 4\,395$. Druhá část výstupu ovšem říká, že jsme získali 4 284 odpovědí (*Responses*), tedy o 111 méně. Rozdíl byl způsoben tím, že ne všichni respondenti respektovali naši žádost o uvedení tří důvodů.

Jak výstup číst? Vidíme, že jsou v něm tři sloupce. N jsou absolutní četnosti odpovědí (*Responses*) – jejich součet je 4 284. Ve druhém sloupci (*Responses – Percent*) jsou **podíly jednotlivých odpovědí vztažených k celkovému počtu odpovědí**. Takže kategorie 1 („Už mám tolik dětí, kolik jsem chtěl/a"), která byla volena 800krát, činí $(800 / 4\,284) \times 100 = 18{,}7$ % ze všech uvedených odpovědí. Ve třetím sloupci (*Percent of Cases*) jsou procentuální **podíly jednotlivých odpovědí z počtu validních respondentů**, to je z 1 465. Takže oněch 800 odpovědí důvodu č. 1 uvedlo 54,6 % respondentů $(800 / 1\,465) \times 100 = 54{,}6$. U tohoto sloupce si všimněme, že součet procent dává dohromady 293 %, což je způsobeno tím, že každý respondent mohl vybrat tři odpovědi (ale ne všichni této možnosti plně využili – pokud by tomu tak bylo, činil by tento součet 300 %).[77] Pro výklad výsledků můžeme použít jak procenta odpovědí, tak i procenta z případů. Pochopitelnější pro čtenáře nebo posluchače, když budeme výsledky tohoto druhu prezentovat, jsou asi výmluvnější procenta z celku odpovědí, neboť ta dávají součet 100 %. Rozhodnutí o tom, která procenta jsou u vícenásobných odpovědí smysluplná, ale vždy záleží na účelu analýzy.

Co lze z výstupu dále vyčíst? Především nás zajímá, které důvody respondenti uváděli nejčastěji. Nejčastěji byl uveden důvod č. 1 („Už mám tolik dětí, kolik jsem chtěl/a" – 19 %), dále důvod č. 6 („Měli bychom vážné existenční potíže" – 12 %) a důvod č. 8 („Příliš se obávám, jaká budoucnost by moje děti čekala" – 11 %). Nejméně častým důvodem byla položka č. 9 („Mít děti dnes nemá žádný význam" – 1 %). Do dalšího meritorního (věcného) výkladu této problematiky se zde nebudeme pouštět a ponecháme to na čtenářích.

Závěrem poznamenejme, že velice časté jsou vícenásobné odpovědi u otevřených otázek (proč asi?) a velice časté je i zaškrtávání jednotlivých odpovědí, které vyústí v několik dichotomických proměnných.

[77] Údaj 293 % lze také po vydělení stem interpretovat jako průměrný počet odpovědí, které poskytl jeden respondent. V našem příkladu tedy jeden respondent poskytl průměrně 2,93 odpovědi ze tří možných (tedy téměř všichni respondenti byli vzorní).

3.4 Rozložení spojitých proměnných

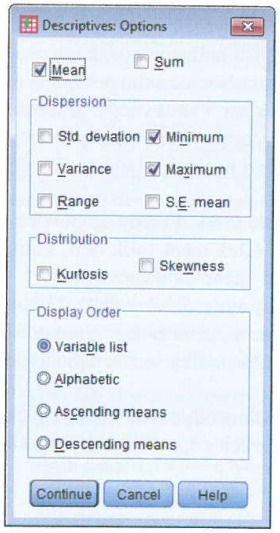
3.4.1 Kontrola nekategorizovaných proměnných

Nekategorizované (kardinální) proměnné mají často tu vlastnost, že mají velký rozsah hodnot (velké množství variant). Při jejich kontrole a čištění tudíž nemá cenu použít proceduru *Frequencies* – dostali bychom totiž příliš mnoho řádků. Namísto *Frequencies* proto použijeme procedury *Descriptives*: *Analyze – Descriptive Statistics – Descriptives*.

Příklad 3.5

V našem datovém souboru z výsledků přijímacího řízení (viz soubor „Vysl-zkousky") je proměnná *scio_osp* obsahující výsledky z testu obecných studijních předpokladů. Víme, že rozsah jejích hodnot se pohybuje v intervalu <0; 100>, je to tedy typická nekategorizovaná proměnná. Zkontrolujme, zdali jsme se při nahrávání jejích hodnot nedopustili nějakých překlepů.

V okně *Descriptives* vybereme proměnnou *scio_osp* a kliknutím na šipku ji vložíme do okénka *Variable(s)*. Pak klikneme na tlačítko *Options* (viz obr. 3.7) a v dialogovém okně si zaklikneme požadavek na minimální a maximální hodnotu a ještě na průměr (*Mean*). Poté klikneme na *Continue* a pak na *OK*.



Obr. 3.7 Zadávání příkazu pro kontrolu kardinálních proměnných

Ve výstupu 3.5 se objeví následující tabulka:

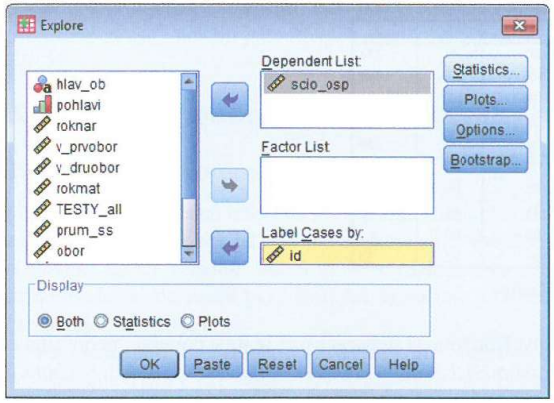
Descriptive Statistics

	SCIO_OSP	Valid N (listwise)
N	180	180
Minimum	7	
Maximum	772	
Mean	78,44	

Výstup 3.5

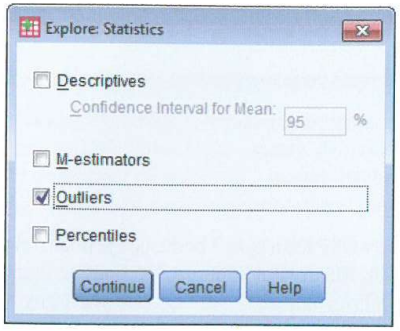
Tabulka říká, že minimální hodnota skóre v OSP testu byla 7 bodů, což je podezřele nízká hodnota a měli bychom ji zkontrolovat. Maximální hodnota 772 bodů je jasný omyl, vždyť nejvyšší možný percentil je 100. Průměr je 78,84, což naznačuje, že chybných údajů s hodnotou nad 100 není sice v datech příliš mnoho, ale každopádně je třeba celé rozložení zkontrolovat. Tím jsme provedli první krok čištění dat a musíme postoupit ke kroku druhému.[78] Máme-li spojitou proměnnou (jako např. *scio_osp*), nevíme přesně, jakou hodnotu máme hledat. Víme sice, že přinejmenším dvě hodnoty jsou pochybné: 7 a 772, ale nevíme, jestli tam nejsou ještě další chyby. Abychom je nalezli, použijeme proceduru *Explore*: *Analyze – Descriptive Statistics – Explore*.

Do okénka *Dependent List* zadáme proměnnou, kterou chceme kontrolovat (*scio_osp*), a do okénka *Label Cases by* vepíšeme identifikační proměnnou. Klikneme na tlačítko *Statistics* a zvolíme *Outliers* (viz obrázky 3.8a, 3.8b níže).



Obr. 3.8a Zadávání příkazu pro vyhledání problematických hodnot proměnné

[78] Upozorněme na tomto místě, že zmíněný postup hledání chyb musíme provést u všech proměnných v datovém souboru. Znovu opakujeme (opakování je totiž matka moudrosti), že důkladná kontrola proměnných je základem úspěšné a kvalitní analýzy.



Obr. 3.8b Zadávání příkazu pro vyhledání problematických hodnot proměnné

Po klinutí na *Continue* a *OK* získáme výstup 3.6. V něm jsou důležité poslední dva sloupečky nadepsané *ID* a *Value*. Sloupec *Value* udává pět nejvyšších hodnot proměnné, které se v souboru vyskytují (v horní polovině tabulky nad čarou, která je označena jako *Highest*), a dále pět nejnižších hodnot dané proměnné (pod čarou v části *Lowest*). Vidíme v něm, že hodnotu 772 měl uchazeč číslo 30 (viz sloupec ID),[79] hodnotu 645 uchazeč č. 150 a hodnotu 181 uchazeč č. 180.

Extreme Values

			Case Number	ID	Value
SCIO_OSP	Highest	1	30	30	772
		2	150	150	645
		3	180	180	181
		4	5	5	93
		5	145	145	86
	Lowest	1	177	177	7
		2	63	63	52
		3	44	44	55
		4	90	90	57
		5	97	97	58

Výstup 3.6 Výstup z procedury *Explore*

To jsou zřetelné překlepy. Hodnota 93 uchazeče č. 5 je již v pořádku, neboť maximální výsledek byl 100. U nejnižších hodnot je hodnota 7 podezřelá a měli bychom ji zkontrolovat. Hodnota 52 je již očividně v pořádku.

[79] To, že se údaj ve sloupci *ID* shoduje s údajem ve sloupci *Case Number* (což je řádek datové matice), je v tomto případě náhoda. Nemělo by nás to vést k domněnce, že nahrávat identifikační číslo respondenta či případu (ID) je zbytečné. Ne, není to zbytečné a každá datová matice SPSS by jako první proměnnou měla mít právě ID.

V proměnné *scio_osp* jsme tedy detektovali celkem čtyři chyby, které musíme opravit způsobem popsaným v předchozím oddíle – výhodou zde je, že už nemusíme vyhledávat jejich identifikace v datové matici, neboť to za nás udělala procedura *Explore* (a *Outliers*).

3.4.2 Popis rozložení kardinální proměnné

V případě, kdy sledovaná proměnná je proměnnou ordinální s mnoha variantami nebo když se jedná o proměnnou intervalovou, třídění prostřednictvím procedury *Frequencies* nemá smysl. Proto se také obvykle v tomto případě nehovoří o četnosti určité hodnoty, neboť hodnoty kardinální proměnné mají malé četnosti (stejná hodnota se v souboru neopakuje příliš často).[80] A připomínáme, že adekvátním grafickým zobrazením rozložení kardinální proměnné není sloupcový graf, ale **histogram**.

Kardinální proměnné proto většinou netabelujeme (nevytváříme tabulku rozložení četností), ale tendenci v datech vyjadřujeme prostřednictvím sumarizujících čísel neboli statistických charakteristik. Jejich výhodou je, že prostřednictvím několika čísel (charakteristik) ilustrují základní vlastnosti rozložení. Srovnáváním těchto charakteristik u různých proměnných porozumíme tomu, co se v daných proměnných děje a jak se navzájem odlišují nebo jak jsou si podobné. K těmto sumarizujícím číslům patří **charakteristiky polohy** (střední hodnoty) a **charakteristiky rozptýlenosti** (variability). Jak střední hodnoty, tak míry rozptýlenosti ale umíme stanovit ne pouze pro proměnné kardinální, ale i pro proměnné nominální a ordinální. Na typu proměnné závisí i použití jednotlivých charakteristik.

3.5 Střední hodnoty a míry variability

3.5.1 Nominální proměnné

U nominální proměnné můžeme určit pouze jednu charakteristiku ze středních hodnot, a to **modus** (*mode*). Modus je kategorie s největší četností, tedy kategorie, která obsahuje nejvíce případů. Stává se, že v rozložení kategorií proměnné není pouze jedna modální, ale mohu být i dvě (pak hovoříme o bimodálním rozložení) nebo tři (trimodální rozložení).

Při zkoumání, jak jsou jednotlivé kategorie obsazeny, tedy do jaké míry variují, se zajímáme o **míry variability**. Vycházíme přitom z konceptu **koncentrace** – sledujeme, zdali některá nebo některé kategorie na sebe váží více jednotek než jiné (jsou více „naloženy"). Pokud je koncentrace nízká, takže jednotlivé kategorie jsou obsazeny víceméně rovnoměrně, jsou data hodně rozptýlena a příslušné míry variability

[80] Je ovšem pravda, že i spojitý znak lze zobrazit v tabulce rozložení četností, ale pouze tehdy, když z této proměnné vytvoříme proměnnou kategorizovanou tak, že stanovíme intervaly hodnot (např. příjmové skupiny, věkové skupiny, intervaly výsledku testu OSP).

pro nominální proměnnou budou nabývat vysokých hodnot. Jestliže vypočtená míra variability je rovna nule, pak jsou kategorie proměnné nulově rozptýleny, data jsou tedy koncentrována pouze do jedné kategorie, takže jsou plně homogenní. Platí proto, že čím vyšší je hodnota, která charakterizuje variabilitu, tím jsou data méně koncentrována a tím vyšší je také heterogenita souboru, a kategorie proměnné jsou z hlediska počtu případů, které obsahují, naloženy víceméně podobně. Míry koncentrace (nebo variability, chcete-li) jsou u nominální proměnné tyto:

- **Variační poměr** $v = 1 - n_{Mo} / n$, (3.1)
  kde n_{Mo} je četnost modální kategorie a n je velikost souboru.

- **Nominální rozptyl** (variance, zvaný též Giniho odchylka)
  $nomvar = \Sigma ((p_i \times (1 - p_i))$, (3.2)
  kde p_i jsou relativní četnosti jednotlivých kategorií (pozor, ne procenta, tedy nenásobená stem) a řecký symbol Σ říká, že *nomvar* vznikne jako součet všech jednotlivých výpočtů, v nichž je každá relativní četnost násobená toutéž relativní četností odečtená od jedné (což je slovní přepis operací uvedených v závorkách).

Míry rozptýlenosti (variability) jsou pro větší přehlednost vyjadřovány ve standardizované podobě, což značí, že nabývají hodnot z intervalu <0; 1>. Standardizace dosáhneme tím, že hodnotu nominálního rozptylu dělíme počtem kategorií sledované proměnné. Pak hovoříme o normalizovaném nominálním rozptylu:

- **Normalizovaný nominální rozptyl** (variance)
  $norm.nomvar = K \times \text{nomvar} / (K-1)$, (3.3)
  kde K je počet kategorií nominální proměnné. Abychom mohli tuto charakteristiku vypočítat, musíme znát nejdříve *nomvar*. Hodnoty *norm.nomvaru* se pohybují v rozmezí od 0 do 1. Čím více se tato hodnota blíží k jedné, tím méně jsou data koncentrována a jsou rovnoměrněji rozložena do jednotlivých kategorií. Je-li hodnota rovna jedné, pak jsou všechny kategorie obsazeny stejným počtem případů.[81]

Míry variability mají pochopitelně smysl pouze tehdy, když srovnáváme několik nominálních proměnných, pro jednu proměnnou nemají tyto údaje žádný smysl.[82]

Příklad 3.6

Ve výzkumu EVS 1999 (soubor „EVS99-cvicny") byla respondentům položena otázka, proč si myslí, že u nás lidé žijí v nouzi. Určeme modus jako střední charakteristiku a míry variability. Rozložení této proměnné (*q11*) ukazuje výstup 3.8.

[81] Podrobněji k nominálnímu a normalizovanému nominálnímu rozptylu v knize Řehák a Řeháková (1986).

[82] Obdobně smysluplné je srovnávat variabilitu (rozptýlenost) u stejné proměnné, ale u různých skupin (muži vs. ženy, mladší vs. starší apod.).

q11 Proč lidé žijí v nouzi - 1. důvod

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 mají smůlu	285	14,9	15,6	15,6
	2 jsou líní	786	41,2	43,0	58,6
	3 je to bezpráví	341	17,9	18,7	77,3
	4 součást pokroku	322	16,9	17,6	94,9
	5 nic z uvedeného	93	4,9	5,1	100,0
	Total	**1827**	**95,8**	**100,0**	
Missing	-2 neodpověděl/a	25	1,3		
	-1 neví	55	2,9		
	Total	**81**	**4,2**		
Total		**1908**	**100,0**		

Výstup 3.8 Rozložení proměnné „proč lidé žijí v nouzi" (q11)

a) Modální kategorií je kategorie **2** (lidé žijí v nouzi, protože *jsou líní*)

b) variační poměr $v = 1 - 786/1827$ (pozor, do n dosazujeme pouze součet respondentů s platnými odpověďmi[83]) $= 1 - 0{,}43 =$ **0,57** (viz vzorec 3.1)

c) *nomvar* = **0,722** (viz vzorec 3.2)

Pozn. Pro výpočet *nomvaru* musíme udělat několik ručních výpočtů, výhodné je použít tabulkový procesor Excel. Tuto excelovskou tabulku (viz níže), v níž jsou již jednotlivé výpočty předdefinovány, přikládáme jako soubor *vyp_nomvar.xls*, který je uložen na přiloženém CD.

q11 Proč u nás lidé žijí v nouzi

i	p	p*(1-p)	
1	0,156	0,132	
2	0,430	0,245	
3	0,187	0,152	
4	0,176	0,145	
5	0,051	0,048	
Součet	1,00	**0,722**	= nomvar

d) $norm.nomvar = 5 \times 0{,}722 / (5 - 1) = 3{,}61 / 4 =$ **0,903** (viz vzorec 3.3)

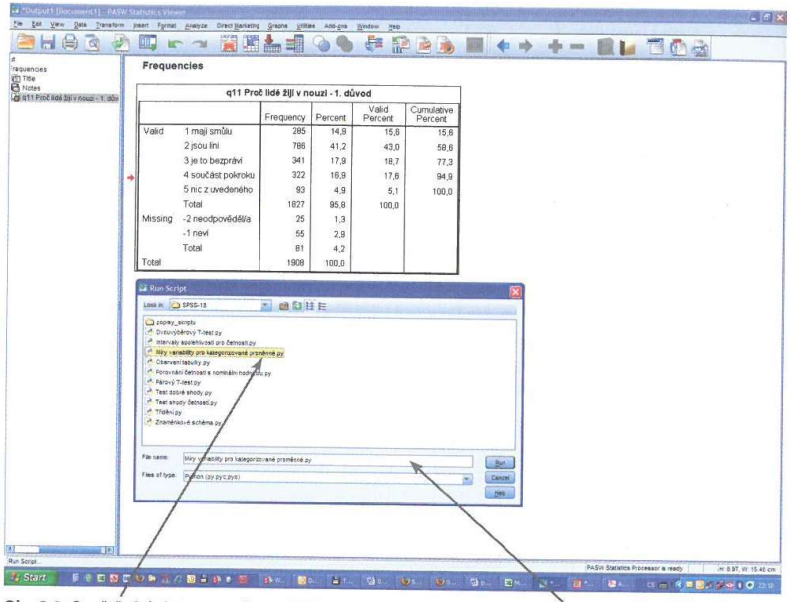
Z vypočtených údajů vyplývá, že míra koncentrace není vysoká, data jsou rozptýlena do všech kategorií.

Nyní, když je nám logika výpočtu jasná, si můžeme dovolit uplatnit velké zjednodušení. Kolegové z české pobočky zastupující IBM SPSS pod vedením doc. Řeháka zpracovali pro některé časté statistické výpočty, jež software SPSS „neumí", speciální malé prográmky, jimž se říká **skripty**. Ty se nasazují v prostředí SPSS na jeho výstupy a požadované charakteristiky okamžitě vypočítají, takže nemusíme nic ani ručně, ani v Excelu počítat. My je budeme postupně představovat, abychom se s nimi naučili pracovat. Tyto prográmky jsou jako zvláštní soubory součástí učebních materiálů, postupně si je s příslušnými lekcemi stahujte a ukládejte si je do vašeho počítače, nejlépe do zvláštního adresáře, vynalézavě nazvaného např. „Skripty".[84]

[83] Při řešení tohoto vzorce nejdříve provedeme dělení čísel a výsledek odečteme od jedné.

[84] Skripty jsou i součástí přiloženého CD, jsou ve verzi pro SPSS 17, 18, 19 a 22. Skripty je možné také stáhnout z webových stránek společnosti Acrea (http://www.acrea.cz/centrum-vyuky), která se v ČR stará o program SPSS a o jeho distribuci.

Skript máme i na míry variability nominální proměnné. Spustíme jej následovně: Necháme si udělat nám již známé rozložení četností proměnné *q11*: *Analyze – Descriptive Statistics – Frequencies*. Ve výstupu označíme tabulku kliknutím levým tlačítkem myši. U tabulky se objeví tučná červená šipka (viz obr. 3.9).



Obr. 3.9 Spuštění skriptu pro míry variability (a nalezení souborů typu Python)

Tím máme tabulku připravenou pro další výpočty – pustíme na ni skript, který se jmenuje *Míry variability pro kategorizované proměnné.py*. Jak to provedeme? Klikneme na tlačítko *Utilities* a poté na tlačítko *Run Script*. Objeví se dialogové okno, v němž zvolíme skript (musíte ale počítači říci, kde ho má ve vašem počítači hledat, tj. musíte mu popsat cestu, kde máte na svém počítači tento skript pro výpočet variability uložen – viz obr. 3.9). Pokud se vám v dialogovém okně *Run Script* neobjeví nabídka souborů skriptů, vepište do posledního řádku, do rámečku *Files of Type* (typ souborů) název *Python* (řádek vám tuto možnost nabídne). Kliknutím na *Run* programek spustíme.

Před samotným výpočtem se počítač ještě zeptá, jaké charakteristiky variability chceme vypočítat. Jelikož je naše proměnná nominální, budeme požadovat variační poměr, *nomvar* a *norm.nomvar* (viz obr. 3.10). Vypočtené míry variability jsou ve výstupu 3.9.

Obr. 3.10 Volba charakteristik variability

Míry variability - q11 Proč lidé žijí v nouzi - 1. důvod

Variační poměr	,570
Nominální variance	,722
Norm. nominální variance	,903

Četnosti vstupující do výpočtu: 285,0 786,0 341,0 322,0 93,0
Počet platných případů: 1827,0

Výstup 3.9 Vypočtené míry variability

Zkontrolujme tento výsledek s našimi ručními výpočty. Jsou v pořádku. Variační poměr je 0,57, nominální variance (rozptyl – *nomvar*) je zde 0,72, my jsme vypočítali 0,722. Normalizovaný nominální variance (rozptyl – *norm.nomvar*) je zde 0,90, my jsme vypočítali 0,903.

3.5.2 Ordinální proměnné

U ordinálních proměnných[85] můžeme jako údaj o střední hodnotě použít modální kategorii (modus), ale výhodnější je použít **medián** (*median*). Medián je hodnota, která dělí rozložení souboru seřazeného podle hodnot této proměnné na dvě poloviny. 50 % jednotek má hodnotu nižší než je medián a 50 % má hodnotu vyšší než je medián. Ordinální proměnné, s nimiž se často pracuje v sociologických analýzách, mají většinou malý počet kategorií (variant), a proto se v takové situaci určuje **mediánová kategorie**. Je to taková kategorie, která splňuje podmínku, že její kumulativní četnost v sobě zahrnuje minimálně 50 % případů (tedy 50 % hodnot v souboru je menších nebo stejných než mediánová kategorie).

Medián patří do kategorie tzv. **kvantilů** a my si o nich povíme více za chvíli, až budeme hovořit o středních hodnotách pro kardinální znaky. U nich mají totiž kvantily velké a smysluplné využití.

[85] Pro hlubší vhled do problematiky ordinálních proměnných a do možností jejich deskripce doporučujeme pročíst stať Jana Řeháka (1976). Velmi inspirativní jsou pasáže o diskrétních (rozpojitých) a kontinuálních (spojitých) typech ordinálních vlastností.

Míry variability pro ordinální proměnné jsou:

- **Variační rozpětí**, což je rozdíl mezi maximální a minimální hodnotou znaku.
- **Ordinální rozptyl (variance)** $dorvar = 2 \times \Sigma ((P_i \times (1 - P_i))$, (3.4)
  kde P_i jsou relativní kumulativní četnosti. Počítá se de facto stejně jako *nomvar*, pouze s tím rozdílem, že pracujeme s relativními kumulativními četnostmi a výsledný součet násobíme dvěma.
- **Normalizovaný ordinální rozptyl** $norm.dorvar = 2 \times dorvar / (K - 1)$, (3.5)
  kde K je počet kategorií ordinální proměnné.[86]

Příklad 3.7

Respondenti ve výzkumu EVS 1999 vyjadřovali svůj postoj k výroku „pracovat je povinnost". Výsledky uvádí výstup 3.10. Mediánovou kategorií je varianta 2 (souhlasí), jako u první v ní kumulativní četnost dosahuje 50 % respondentů. Mimochodem je to současně i kategorie modální.

q17_4 Pracovat je povinnost

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 rozhodně souhlasí	358	18,8	19,0	19,0
	2 souhlasí	831	43,6	44,0	62,9
	3 ani souhlas ani nesouhlas	368	19,3	19,5	82,4
	4 nesouhlasí	278	14,6	14,7	97,1
	5 rozhodně souhlasí	55	2,9	2,9	100,0
	Total	1889	99,0	100,0	
Missing	-2 neodpověděl/a	5	,2		
	-1 neví	14	,7		
	Total	19	1,0		
Total		1908	100,0		

Výstup 3.10 Rozložení proměnné „pracovat je povinnost" (q17_4)

Dorvar = 2 × 0,56 = 1,121 (podle rovnice 3.4 a pomocných výpočtů v tabulce níže).[87]

q17_4 Pracovat je povinnost

i	p	p*(1-p)		
1	0,190	0,154		
2	0,629	0,233		
3	0,824	0,145		
4	0,971	0,028		
5	1,000	0,000		
Součet	3,61	**0,560**	**1,121**	**dorvar**

norm.dorvar = 2 × 1,121 / (5 − 1) = 2,242 / 4 = **0,561** (podle rovnice 3.5)

[86] Podrobněji k ordinálnímu a normalizovanému ordinálnímu rozptylu v knize Řehák a Řeháková (1986).

[87] Excelovský soubor s tímto výpočtem je přiložen na CD.

Pokud použijeme příslušný skript (viz obr. 3.10) a navolíme charakteristiky pro ordinální proměnnou, dostaneme tytéž výsledky, jaké jsme my získali ručním výpočtem.

3.5.3 Kardinální proměnné

U kardinálních znaků lze jako charakteristiky střední polohy použít jak modus, tak i medián. Medián zde určujeme tak, že u souborů, které mají lichý počet prvků, je hodnota mediánu rovna hodnotě středního prvku při seřazení hodnot od nejmenší po největší. Při sudém počtu prvků se medián počítá jako aritmetický průměr hodnot dvou středních prvků.

Speciální střední hodnotou pro kardinální proměnné je (všem dobře známý) **aritmetický průměr** (*mean*)

$$\bar{x} = \frac{\sum x}{n}. \quad (3.6)$$

Vypočteme jej tak ($\bar{x}$ čteme jako „x s pruhem"), že sečteme všechny hodnoty v souboru a podělíme velikostí souboru. Ačkoliv se průměr velmi často při prezentaci nějaké kardinální proměnné používá (od statistiků se např. dozvídáme, jaká byla v Česku průměrná měsíční mzda v roce 2011, jaký byl u nás průměrný počet litrů vypitých piv – zde jsme „nejlepší na světě" –, kolik spotřebujeme průměrně kilogramů zeleniny za rok – zde naopak k premiantům vůbec nepatříme – atd.), není v mnoha případech úplně tou nejvhodnější charakteristikou. Je totiž ovlivnitelný odlehlými hodnotami – je na ně citlivý.

Střední hodnoty jakožto míry centrální tendence (modus pro nominální proměnné, mediánová kategorie pro ordinální proměnné a aritmetický průměr pro kardinální proměnné) nebývají často pro rozložení dostačující charakteristikou, a proto je vhodné uvádět spolu s nimi i statistické charakteristiky rozptýlení neboli míry variability (rozptýlenosti).[88]

Míry variability pro kardinální proměnné jsou:

- **Rozptyl** (*variance*), značený symbolem s^2 nebo též *var x*, patří k základním pojmům statistiky a všichni, kdo se budou ve statistické analýze pohybovat, se s ním budou často setkávat. Popišme si proto slovně, jak se rozptyl vypočítává, abychom mu dobře a na věky věků rozuměli. Takže, vypočítá se tak, že od každé hodnoty dané proměnné odečteme její průměr. Získáme tak odchylky od průměru, z nichž některé budou kladné, jiné záporné (bude-li průměr např. 8 a jedna z našich hodnot bude 5, je odchylka 5 − 8 = −3; bude-li hodnota 10, je odchylka 10 − 8 = 2). Všechny takto stanovené odchylky musíme sečíst. Ale jelikož platí, že součet všech

[88] Hendl (2004, s. 95) upozorňuje, že „omezenost středních hodnot spočívá v tom, že udávají pouze to, kolem jaké hodnoty se data centrují, respektive které hodnoty jsou nejčastější, ale data se stejnou střední hodnotou mohou mít různou rozptýlenost".

těchto odchylek od průměru je roven nule, umocníme před sečtením všechny odchylky na druhou (tím se mimo jiné také zdůrazní hodnoty, které leží ve velké vzdálenosti od průměru) a teprve poté je sečteme. Aby tento součet nebyl ovlivněn počtem měření (čím více měření budeme mít, tím více hodnot bude mít daný znak a tím vyšší bude i výsledný součet), musíme jej standardizovat tím, že součet umocněných odchylek od průměru vydělíme celkovým počtem hodnot znaku. Matematicky zapsáno vše vypadá následovně:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}. \qquad (3.7)$$

- **Směrodatná odchylka** (*standard deviation*), značená s nebo σ (sigma), je druhou odmocninou rozptylu. Používá se častěji než rozptyl, a to z toho důvodu, že eliminuje jednu velkou nevýhodu rozptylu, která spočívá v tom, že při jeho výpočtu umocňujeme jednotky na druhou, což jim ubírá smysluplnou interpretaci – když budeme mít průměrný výdělek v korunách, bude rozptyl v korunách na druhou, u vypitých piv to budou piva na druhou (možná sen některých českých konzumentů, ale v realitě neexistující jednotka). Naproti tomu směrodatná odchylka jakožto druhá odmocnina rozptylu vrací hodnoty proměnné do původních jednotek, v nichž byla měřena, čímž nazpět získává srozumitelnou interpretaci.

$$s = \sqrt{s^2} \qquad (3.8)$$

- **Variační koeficient** V je velmi užitečnou charakteristikou variability. Vypočítá se jako podíl rozptylu k průměru a obvykle se násobí stem.[89]

$$V = \frac{s}{\bar{x}} \times 100 \qquad (3.9)$$

Nyní několik poznámek k právě představeným charakteristikám střední hodnoty (polohy) a variability. Průměr bychom měli používat pouze tehdy, když jsou hodnoty proměnné přibližně symetricky rozloženy kolem jednoho vrcholu (to zjistíme „okometricky" pohledem na histogram četností). Proměnné mohou mít stejný průměr, ale jejich rozptyl může být odlišný, takže jejich směrodatná odchylka je různá. Malá směrodatná odchylka je znamením, že průměr je dobrým a vhodným popisem dané proměnné. Velká směrodatná odchylka vždy naznačuje, že data pocházejí ze souboru velmi heterogenních jednotek, což dále značí, že používání průměru pro popis proměnné není smysluplné. Výrazy „malá" a „velká" směrodatná odchylka

[89] Výhoda variačního koeficientu je, že v situacích, kdy srovnáváme rozptýlenost údajů, které jsou měřené v různých jednotkách, je variační koeficient smysluplný a směrodatná odchylka nikoliv (například některé země uvádějí roční příjmy a jiné měsíční, jindy jsou příjmy v různých měnách apod.).

jsou ovšem relativní, závisí na jednotce měření a také na kontextu. Šedesátivteřinová směrodatná odchylka od průměrného času maratonce *XY* v jeho 10 maratonských bězích za poslední tři roky je jistě malá směrodatná odchylka (nejlepší běžci dnes běhají maraton přibližně za 2 hodiny a 5 minut), zatímco pětivteřinová směrodatná odchylka od průměrného času sprintera *AB* v běhu na 100 metrů (tuto vzdálenost uběhnou světoví sprinteři za 10 vteřin) za poslední tři roky by byla obrovská.

Rozptyl a směrodatná odchylka jsou podobně jako průměr citlivé na extrémně odlišné hodnoty. Několik extrémních hodnot může velmi zvýšit velikost směrodatné odchylky. Například velikost směrodatné odchylky 20 u průměru 150 říká, že velká část hodnot této proměnné leží 20 jednotek na každou stranu od průměru, takže se pohybují v intervalu 130 až 170 (jak velká to je část, si rozebereme v následující kapitole). Směrodatná odchylka je většinou různá od nuly, nule je rovna pouze v případě, kdy všechny hodnoty proměnné jsou shodné, a tedy konstantní – pak proměnná není de facto proměnná, ale konstanta.

Variační koeficient je dobrým nástrojem na odhad míry homogenity či heterogenity souboru. Velmi hrubé pravidlo říká, že pokud je variační koeficient vyšší než 50 %, pak je to signál, že statistický soubor jednotek je v této proměnné natolik nesourodý, že použití statistického průměru je již neospravedlnitelné (Swoboda, 1977).

student	ZK body	(x_i - prům.)	(xi - prům.)^2
x1	24	-9,4	88,7
x2	34	0,6	0,3
x3	34	0,6	0,3
x4	32	-1,4	2,0
x5	36	2,6	6,7
x6	31	-2,4	5,9
x7	34	0,6	0,3
x8	32	-1,4	2,0
x9	37	3,6	12,8
x10	33	-0,4	0,2
x11	41	7,6	57,5
x12	40	6,6	43,3
x13	43	9,6	91,8
x14	47	13,6	184,4
x15	42	8,6	73,6
x16	25	-8,4	70,9
x17	35	1,6	2,5
x18	30	-3,4	11,7
x19	26	-7,4	55,1
x20	33	-0,4	0,2
x21	43	9,6	91,8
x22	36	2,6	6,7
x23	31	-2,4	5,9
x24	40	6,6	43,3
x25	18	-15,4	237,8
x26	45	11,6	134,1
x27	47	13,6	184,4
x28	15	-18,4	339,3
x29	14	-19,4	377,1
x30	27	-6,4	41,2
x31	31	-2,4	5,9
N = 31	1036	0	2178
	Σ		Σ

průměr =	1 036 : 31 = 33,42
rozptyl =	2 178 : 31 = 70,2
směrodatná odchylka =	√70,2 = 8,4
variační koeficient =	(8,4 : 33,4)*100 = 25,1

Obr. 3.11 Ukázka výpočtu rozptylu, směrodatné odchylky a variačního koeficientu

Ukázka výpočtu průměru a měr variability je předvedena na obr. 3.11 (data jsou ze souboru „vysl-zkousky", výpočet lze kontrolovat v excelovském souboru „variance--vyp.xls" na CD). Průměrný zisk bodů u zkoušky v souboru 31 studentů byl 33 bodů (rozpětí této proměnné bylo od 0 bodů do maxima 50). Rozptyl bodového výsledku byl 70. Směrodatná odchylka 8 bodů říká, že se většina čísel odchyluje o 8 bodů od průměru v obou směrech, pohybuje se tedy mezi 25 a 41 body. Variační koeficient je 25 %, použití průměru jako kondenzovaného výrazu o statistickém charakteru této proměnné je oprávněné.

Variabilitu proměnné můžeme popsat ještě dalšími užitečnými charakteristikami. Říká se jim obecně **kvantily**. Jelikož se budeme v naší analytické práci v sociálních vědách setkávat především s kvantily, které jsou vyjadřovány v procentech, budeme zde hovořit o **percentilech** (empirických percentilech). My jsme v této kapitole o jednom percentilu de facto již hovořili, a to když jsme představili medián. Medián proměnné *X* dělí počet jednotek souboru na dvě přesně stejné poloviny, 50 % jednotek má hodnotu pod mediánem a 50 % hodnotu vyšší než medián. Z tohoto hlediska je medián 50% percentil.

V praxi statistické analýzy se velmi často pracuje s tzv. **kvartily**, které dělí soubor na čtvrtiny. Stanovením hodnoty prvního neboli dolního kvartilu (Q_I) víme, že 25 % jednotek souboru je pod touto hodnotou. Hodnota druhého kvartilu (Q_{II}) je hodnotou mediánu, velikost třetího neboli horního kvartilu (Q_{III}) určuje, že 75 % souboru je pod touto hodnotou (a samozřejmě 25 % souboru je nad touto hodnotou). Takže když např. zjistíme výpočtem v SPSS, že v testech OSP, to je v testech obecných studijních předpokladů (OSP se pohybují v intervalu 0–100 bodů), byl $Q_I = 61$ bodů, $Q_{II} = 72$ b. a $Q_{III} = 79$ b., pak okamžitě víme, že 25 % účastníků testu mělo bodový výsledek méně než 61 bodů, 50 % účastníků získalo méně než 72 bodů a 75 % účastníků mělo méně než 79 bodů. Když navíc z rozložení dat zjistíme, že nejnižší bodový výsledek byl 24 bodů a nejvyšší 98 bodů, pak také lehce určíme, že 25 % uchazečů, kteří spadli do dolního kvartilu, získalo 24–61 bodů a nejlepších 25 %, kteří byli v horním kvartilu, získalo 79–98 bodů. Dalším způsobem, jak popsat variabilitu znaku, je **mezikvartilové rozpětí** (*interquartile range*), což je rozdíl mezi hodnotou horního (Q_{III}) a dolního (Q_I) kvartilu.

Kromě kvartilů pracujeme někdy též s kvintily, které dělí soubor na pětiny po 20 %, a decily, které rozdělují souboru na desetiny.[90] Decily lze například využít při zkoumání chudoby. U rozložení příjmů nás musí zajímat, jaká je hodnota spodního decilu (to je těch nejchudších) a horního decilu (kolik vydělávají ti nejbohatší). Kromě toho nás také může zajímat, jaké jsou jejich typické sociální charakteristiky.

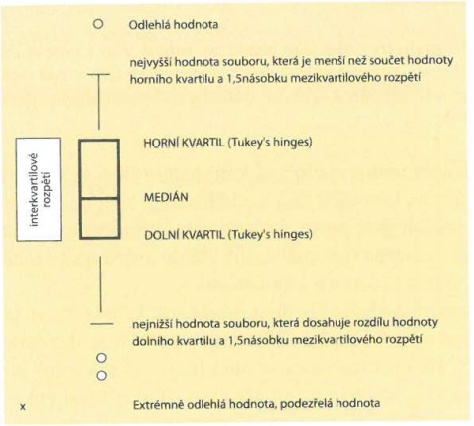
[90] Jen pro upřesnění, ale aby nás to nemátlo. Kvartily nejsou čtyři, ale tři; kvintilů není pět, ale jsou čtyři a decilů není deset, ale pouze devět. Je to jako při dělení úsečky na 4 stejně dlouhé úseky – stačí vám k tomu pouze 3 značky.

Velmi dobrým popisem centrální tendence a rozložení proměnné je tzv. **pětičíselné shrnutí** (*five-number summary*, viz Tukey, 1977) nebo též popis dat pomocí pěti hodnot.[91] Těmito pěti hodnotami jsou, symbolicky zapsáno: **Min – Q_I – Me – Q_{III} – Max**. Řečeno slovně, minimální hodnota proměnné, dolní kvartil, medián, horní kvartil, maximální hodnota proměnné. Vrátíme-li se k našemu příkladu o výsledcích v testu OSP, pak těchto pět hodnot má následující podobu (viz tab. 3.2). O slovní výklad této číselné sumarizace se pokuste sami, inspirovat se můžete našimi předchozími výroky uvedenými v odstavci o percentilech.

Min	Q_I	Medián	Q_{III}	Max
24	61	72	79	98

Tab. 3.2 Pětičíselné shrnutí výsledku v testech OSP

Popis dat pomocí pěti hodnot slouží k sestrojení velmi zajímavého a pro analytické účely značně užitečného grafu. Říká se mu **krabičkový graf** (*Box and Whiskers Graph*). Jeho autorem je americký matematik John Wilder Tukey.[92] Vypadá takto (viz obr. 3.12):

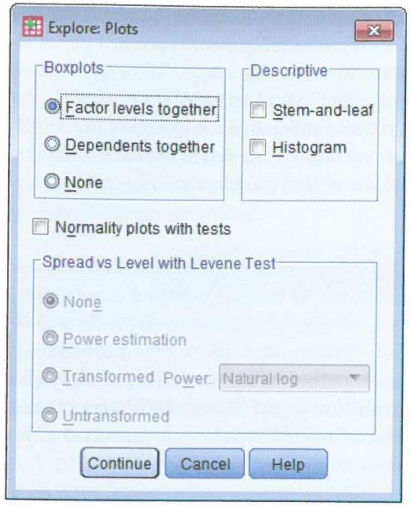


Obr. 3.12 Krabičkový graf

Tento druh grafu získáme v proceduře *Analyze – Descriptive Statistics – Explore*. V dialogovém okně v oddíle *Display* zaškrtneme *plots* (tedy že chceme graf) a pak klikneme na tlačítko *Plots*. V oddíle *Boxplots* (krabičkový graf) zaškrtneme volbu *Factor levels together* (viz obr. 3.13).

[91] Popis lze najít např. u Hendla (2004, s. 101).

[92] Proto se také hodnotám horního a dolního kvartilu v angličtině říká *Tukey's hinges* neboli Tukeyho stěžejní body.



Obr. 3.13 Dialogové okno pro volby krabičkového grafu (v proceduře *Explore*)

Pozn: Všimněme si, že vedle krabičkového grafu lze zadat také histogram (stejně jako v proceduře *Frequencies* nebo v proceduře *Graphs*), dále graf „stonek a lodyha" (*stem-and-leaf*), kterým se zde však nebudeme zabývat, protože pro nás nemá větší význam, a konečně grafy testující normalitu rozložení – k jejich užití se dostaneme v příští kapitole.

A ještě jeden typ měr, který charakterizuje rozložení kardinální (ale i ordinální) proměnné, si uveďme. Je sice již trochu specifičtější a v publikacích sociálních věd se s ním příliš často nesetkáváme (zčásti jistě proto, že kardinální proměnné nejsou častou součástí datových souborů sociálněvědních analytiků), ale do kurzu o statistice a SPSS je nutné jej zahrnout. Jsou to míry šikmosti a špičatosti.

Šikmost (*skewness*) je míra symetrie rozložení hodnot proměnné. Lépe řečeno, je to míra jeho asymetrie – ve srovnání s normálním rozdělením, o kterém se dozvíme více v následující kapitole –, neboť šikmost rovnající se nule (nebo blízká nule) indikuje symetrické rozložení, kdy modus, medián a průměr mají shodné nebo velmi podobné hodnoty. Nabývá-li šikmost kladných hodnot, je rozložení zešikmené doprava neboli pravá strana rozložení má delší konec než strana levá. Nabývá-li hodnot záporných, je rozložení zešikmené doleva, jeho levý konec je delší než pravý.

Špičatost (*kurtosis*) je míra indikující, zdali je rozložení špičaté nebo ploché. Čím je rozdělení špičatější, tím více jsou hodnoty soustředěny kolem jeho středu, čím je méně špičaté, tedy plošší, tím častěji obsahuje hodnoty vzdálené od tohoto středu. Je-li koeficient špičatosti vyšší než nula, je rozložení plošší (placatější), je-li menší než nula, je rozložení špičatější než normální rozdělení.

3.6 Výpočty středních hodnot a variability v SPSS

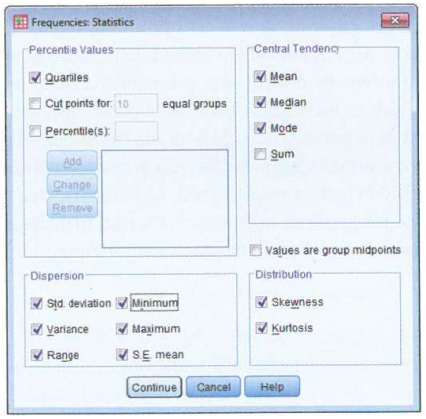
K výpočtům všech výše uvedených charakteristik centrální tendence a variability můžeme v SPSS využít tří procedur: *Frequencies*, *Descriptives* a *Explore*. Ukažme si je postupně všechny při aplikaci na jeden příklad.

3.6.1 Procedura *Frequencies*

Příklad 3.8

Na základě přijímacích zkoušek bylo na FSS MU přijato do bakalářského prezenčního studia celkem 180 studentů. Proveďme zevrubnou analýzu jejich bodového zisku. Pracujeme se souborem „fiktivni.sav" a s proměnnou *testyall*.

Řešení: V proceduře *Analyze – Desriptive Statistics – Frequencies* nebudeme požadovat ve výstupu zobrazení frekvenční tabulky (proto zrušíme zaškrtnutí v okénku *Display frequency table*), naopak klikneme na tlačítko *Statistics* a navolíme výpočty příslušných charakteristik (viz obr. 3.14). SPSS počítá téměř všechny, o nichž jsme na předcházejících stranách hovořili. Všimněme si, že u percentilů nám dává možnost přímo volit kvartily nebo navolit percentily. My jsme se rozhodli, že nám postačují kvartily. Výsledkem našich požadavků je výpočet, který je zobrazen na výstupu 3.11.



Obr. 3.14 Dialogové okno pro volbu výpočtu statistických charakteristik

Statistics

TESTY_all Celkový bodový výsledek v přijímacích testech

N	Valid	180
	Missing	0
Mean		139,71
Std. Error of Mean		,586
Median		140,00
Mode		141[a]
Std. Deviation		7,863
Variance		61,827
Skewness		1,232
Std. Error of Skewness		,181
Kurtosis		6,215
Std. Error of Kurtosis		,360
Range		63
Minimum		120
Maximum		183
Sum		25148
Percentiles	25	136,00
	50	140,00
	75	143,00

a. Multiple modes exist. The smallest value is shown

Výstup 3.11

Co nám výstup 3.11 říká? Nejdříve se v datech musíme zorientovat. Teoreticky mohli uchazeči o studium získat v písemných testech 0–200 bodů (tuto informaci z tabulky nevyčteme, je to danost přijímacího řízení fakulty X). Podívejme se do spodní části tabulky na údaje o dosaženém minimu a maximu. Vidíme, že minimální počet bodů, který přijatí studenti získali, byl 120 a že získaný maximální počet bodů byl 183. V tomto intervalu 120–183 se tedy pohyboval bodový zisk přijatých studentů. Udělejme si sumarizaci podle pěti čísel: Min = 120, Q_I = 136 (tento údaj je úplně na spodku tabulky, v řádku *Percentiles* 25), Medián = 140, Q_{III} = 143 (viz *Percentiules* 75), Max = 183. Průměrné skóre (*Mean*) mělo hodnotu 140 bodů (139,71), nejčastějším bodovým ziskem (*Mode*) bylo 141 bodů. Údaje ze sumarizace podle pěti čísel říkají, že 25 % přijatých získalo mezi 120 a 136 body, dalších 25 % přijatých mělo bodový zisk mezi 136 a 140 body. 75 % uchazečů pak získalo do 143 bodů. Nejlepší čtvrtina uchazečů měla bodový zisk mezi 143 a 183 body.

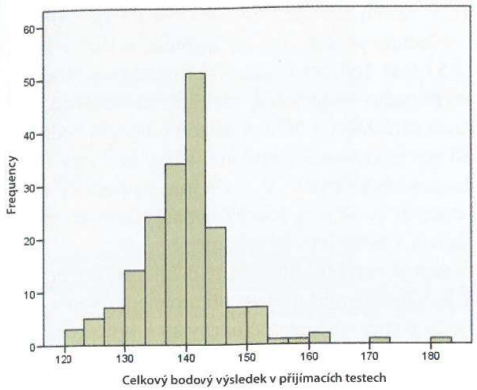
A nyní důležitá poučka. **Údaj o průměru by neměl být nikdy používán osamoceně** bez toho, že bychom jej doplnili informací o variabilitě hodnot znaku.

Rozptyl (*Variance*) je v našem příkladu 61,8, směrodatná odchylka (*Std. Deviation*) je 7,86.[93] Vzhledem k rozsahu stupnice měření (0–200 bodů) to není vysoká směrodatná odchylka a naznačuje, že bodový zisk jednotlivých uchazečů byl poměrně vyrovnaný a že rozptyl v datech nebyl příliš velký.

Dobrým indikátorem toho, jak jsou data rozptýlena, je srovnání průměru, mediánu a modu. Vidíme, že průměr (140), medián (140) i modus (141) se podobají, takže jsou dobrými sumarizujícími ukazateli (jsou dobrým modelem) této proměnné. Pokud se hodnota průměru a mediánu liší výrazně, pak to vždy signalizuje skutečnost, že v datech se vyskytuje nějaká podstatně (možná až extrémně) odlehlá hodnota (nebo několik odlehlých hodnot – tzv. *outliers*). V takovém případě není vhodné pro popis proměnné používat průměru, neboť ten je těmito odlehlými hodnotami ovlivněn, a přednost je třeba dát mediánu.[94] Naše data o přijímacím testu, jak je patrné z obrázku 3.15, jsou rozložena poměrně pravidelně, několik odlehlých hodnot střední tendenci nijak neovlivňuje.

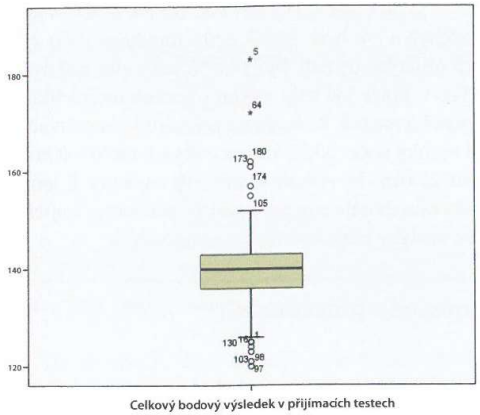
[93] Zkontrolujte, zdali SPSS dělá výpočty správně, a vypočítejte si na kalkulačce druhou odmocninu z hodnoty rozptylu 61,8. Měli byste dostat hodnotu 7,86, tedy hodnotu směrodatné odchylky.

[94] Pro zvídavé dodejme, že klasickým příkladem rozdílu mezi průměrem a mediánem je proměnná příjem. Medián příjmu je vždy menší než průměr. Proč tomu tak asi je? Má smysl používat pro příjem průměr i medián?



Obr. 3.15 Histogram bodového výsledku v přijímacích testech

Tuto informaci dále potvrzuje krabičkový graf s vousy (viz obr. 3.16).



Obr. 3.16 Krabičkový graf bodového výsledku v přijímacích testech

Graf je velmi informativní. Ukazuje, že data jsou těsně rozložena kolem mediánu (tučná čára uprostřed krabičky) a že mezikvartilové rozpětí je úzké (to je ta vertikální délka krabičky, v našem případě je to 7 bodů). V krabičce leží 50 % všech případů. Dolní hrana krabičky je 25. percentil (Q_I) a horní hrana 75. percentil (Q_{III}). Dolní „vousy" (*whiskers*) mají hodnotu 1,5násobku mezikvartilového rozpětí minus hodnotu dolního kvartilu. V našich datech tato hodnota činí 136 – (1,5 × 7) = 125,5. Horní vousy mají naopak hodnotu 1,5násobku mezikvartilového rozpětí plus hodnotu horního kvartilu. To je 143 + (1,5 × 7) = 153,5. V grafu vidíme, že vousy se skutečně pohybují v tomto intervalu. Všechny případy, jejichž hodnota leží pod nebo nad těmito vousy (přesněji řečeno, jejichž hodnota je mezi 1,5- až 3násobkem dolního či horního kvartilu),

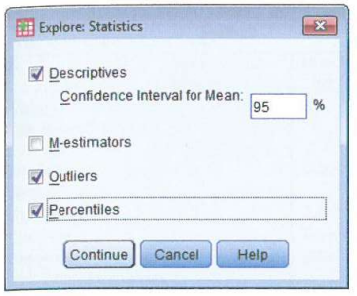
jsou hodnotami odlehlými (*outliers*). V našem případě je 1,5násobek interkvartilového rozpětí 10,5 a trojnásobek je 21. Všechny případy, jejichž hodnota je tedy v intervalu 125,5 až 115 nebo v intervalu 153,5 až 164, jsou hodnotami odlehlými. V grafu jsou znázorněny symbolem ∘ s číslem případu – vidíme tedy, že např. student/ka č. 180 má vysokou odlehlou hodnotu, zatímco student/ka č. 97 má nízkou odlehlou hodnotu. Hodnoty, které jsou vyšší nebo nižší než trojnásobek vertikální délky krabičky (tedy interkvartilového rozpětí), jsou hodnotami extrémními. V grafu jsou vyznačeny symbolem ∗. V našich datech to jsou studenti č. 64 a 5, jejichž hodnoty jsou extrémně vysoké – samozřejmě relativně, vzhledem k ostatním výsledkům.

Dalším indikátorem rozptylu v datech je **variační koeficient**, což je jedna z nejlepších měr relativní variability, neboť je výborným nástrojem při srovnání dvou a více souborů. Jak jsme již uvedli dříve, je to poměr směrodatné odchylky k aritmetickému průměru násobený 100 (je nutné ho vypočítat na kalkulačce, SPSS nemá tento výstup zabudován). Variační koeficient v našem příkladu je (7,86 / 139,7) × 100 = 5,6 %. Představme si nyní jiný soubor, např. studenty přijaté na Fakultu sociálních věd UK, kteří by dělali stejné přijímací testy jako uchazeči o studium na FSS v Brně. Jejich výsledek by byl následující: průměrný výkon v testech by byl v Praze jen o něco vyšší, 141,6, ale „pražská" směrodatná odchylka by byla 19,87, tedy mnohem vyšší než v Brně. Variační koeficient pražských přijatých by tedy byl 14,0 %, tedy více než dvojnásobný ve srovnání s Brnem (5,6 %). V Praze byl tedy výkon v testech našich fiktivních přijatých mnohem více heterogenní a možná, že hodnota průměru byla ovlivněna několika málo studenty, kteří získali vysoký počet bodů, zatímco zbytek mohl mít horší výkon než v Brně. Což by znamenalo, že Brno by získalo kvalitnější studenty. K tomu, abychom tuto otázku vyřešili, bychom museli dále srovnat hodnoty pěti sumarizujících čísel a/nebo si udělat některé grafické analýzy rozložení těchto proměnných.

3.6.2 Analýza kardinální proměnné v procedurách *Descriptives* a *Explore*

Výstup z procedury *Descriptives* přináší v podstatě stejný druh informací jako procedura *Frequencies*, takže se jí není třeba nijak zvláště zabývat, zejména je-li nabídka na výpočet statistických charakteristik na první pohled chudší – není zde např. možnost volit si výpočet percentilů. Zadání výpočtu pro kardinální proměnnou je obdobou zadání, s nímž jsme se seznámili v proceduře *Frequencies*. Procedura *Desriptives* ale umí jednu důležitou statistickou operaci, totiž převod hodnot proměnné na standardizované skóry, tzv. z-skóry. Těmi se budeme zabývat v kapitole 4.

Také zadání pro analýzu prostřednictvím *Explore* je obdobou zadání, s nímž jsme se seznámili v proceduře *Frequencies*. I zde si po proklikání přes *Analyze – Desriptive statistics – Explore* (stále pracujeme se souborem „fiktivni") volíme možnosti výpočtů statistických charakteristik (viz obr. 3.17). Výstup z této procedury má dvě části (viz výstup 3.12a a 3.12b), jež obsahují některé nové informace.



Obr. 3.17 Možnosti statistik v proceduře *Explore*

Descriptives

			Statistic	Std. Error
TESTY_all Celkový bodový výsledek v přijímacích testech	Mean		139,71	,586
	95% Confidence Interval for Mean	Lower Bound	138,55	
		Upper Bound	140,87	
	5% Trimmed Mean		139,42	
	Median		140,00	
	Variance		61,827	
	Std. Deviation		7,863	
	Minimum		120	
	Maximum		183	
	Range		63	
	Interquartile Range		7	
	Skewness		1,232	,181
	Kurtosis		6,215	,360

Výstup 3.12a Ukázka výstupu procedury *Explore*

Ve výstupu 3.12a je to např. údaj o intervalu spolehlivosti průměru (*95 % Confidence Interval for Mean*) – k tomu, co to znamená a jaký to má pro nás význam, se dostaneme v kapitole 5. Dále je zde údaj o upravené hodnotě průměru, o tzv. seříznutém průměru (*5 % Trimmed Mean*). Tento se vypočítává tak, že se nebere v úvahu 5 % nejnižších a 5 % nejvyšších hodnot. Tímto krokem se eliminují případné odlehlé hodnoty a průměr má tak – jako souhrnná charakteristika – vyšší vypovídací schopnost. Jelikož v naší tabulce je rozdíl mezi „normálním" průměrem (139,7) a průměrem „seříznutým" (139,4) jen minimální, je to pro nás signálem, že v hodnotách této proměnné není extrémních hodnot příliš mnoho. V tabulce je také hodnota mezikvartilového rozpětí (*Interquartile Range*).

Výstup 3.12b uvádí hodnoty některých percentilů, jakož i hodnoty dolního kvartilu neboli 25. percentilu (136), mediánu neboli 50. percentilu (140) a horního kvartilu neboli 75. percentilu (143) – tyto údaje jsou zvláště ve sloupci *Tukey's Hinges*.

Percentiles

Percentiles	Weighted Average(Definition 1)	Tukey's Hinges
	TESTY_all Celkový bodový výsledek v přijímacích testech	TESTY_all Celkový bodový výsledek v přijímacích testech
5	127,05	
10	130,10	
25	136,00	136,00
50	140,00	140,00
75	143,00	143,00
90	147,00	
95	150,95	

Výstup 3.12b Ukázka výstupu procedury *Explore*

3.6.3 Dodatek: Analýza ordinální proměnné s dlouhou stupnicí

V sociologickém výzkumu nemáme k dispozici data intervalová příliš často, většinu sociálních vlastností totiž neumíme na intervalových stupnicích změřit. Proto často pracujeme s daty ordinálními, u nichž alespoň konstruujeme dlouhé stupnice měření – z hlediska statistiky jde o ordinální spojité (kontinuální) proměnné (viz Řehák, 1976, s. 421). Podívejme se na následující příklad.

Příklad 3.9

Ve výzkumu EVS 1999 byla respondentům položena následující otázka: *Jak důležitý je Bůh ve Vašem životě?* Respondent odpovídal s pomocí karty, na níž byla tato stupnice:

1 2 3 4 5 6 7 8 9 10
vůbec ne důležitý velmi důležitý

Poznámka: Toto je častý způsob měření některých znaků. Tím, že takto měřená ordinální proměnná má mnoho stupňů měření, mění se na proměnnou semi-kardinální, u níž již má smysl používat mnohé ze statistických operací určených pro intervalové proměnné.

Řešení: Výpočet provedeme prostřednictvím procedury *Frequencies – Statistics*. Výsledek je na výstupu 3.13.

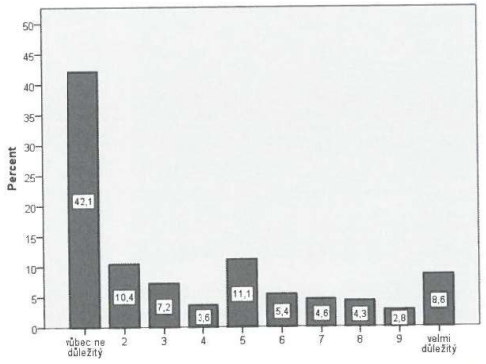
Statistics

Q33 Bůh - důležitost v životě

N	Valid	1846
	Missing	62
Mean		3,63
Std. Error of Mean		,07
Median		2,00
Mode		1
Std. Deviation		3,06
Variance		9,35
Skewness		,858
Std. Error of Skewness		,057
Kurtosis		-,614
Std. Error of Kurtosis		,114
Percentiles	25	1,00
	50	2,00
	75	6,00

Výstup 3.13

Tabulku je vždy dobré doplnit ještě grafem, abychom si učinili představu, jak jsou data rozložena (viz graf na výstup 3.14). Tvar rozložení, opakujeme, je totiž velmi důležitý pro další statistické úvahy (více o tom v další kapitole).



Výstup 3.14 Důležitost Boha v životě jedince v ČR v r. 1999

Zdroj: Data výzkumu EVS 1999.

Jaké poznatky lze z těchto informací získat? Především z grafu vidíme, že procentuální rozložení odpovědí na tuto otázku je velmi nepravidelné a má velmi daleko k rozložení symetrickému. Nejčastější odpovědí byla varianta s kódem 1 „Bůh není v mém životě vůbec důležitý" (42 % respondentů), proto je také kategorií, jak

říká výstup 3.13, modální (modus, *Mode* = 1). Tato informace naznačuje, že značná část české populace není nábožensky založena.[95]

Potvrzují to i další údaje, které vyčteme z výstupu 3.13: hodnota mediánové kategorie (*Median*) 2 říká, že 50 % respondentů nemělo vyšší hodnotu tohoto znaku než 2 a průměrná hodnota (*Mean*) všech respondentů je 3,6. Což znamená, že v průměru není Bůh pro českou populaci příliš důležitý. Směrodatná odchylka je vzhledem k průměru vysoká (3,1), což potvrzuje i vysoká hodnota variačního koeficientu (84,3 %). S průměrem tak nemá příliš smyslu u této proměnné operovat.

Země	Průměr	Směrodatná odchylka	Variační koeficient	N
ČR	3,6	3,1	86	1 846
Dánsko	4,0	2,8	70	1 001
Švédsko	4,1	3,0	73	995
Francie	4,4	3,0	68	1 580
Velká Británie	4,9	3,2	65	960
SRN	5,0	3,1	62	1 988
Slovinsko	5,0	3,2	64	980
Nizozemsko	5,0	3,1	62	999
Bulharsko	5,2	3,2	62	965
Rusko	5,3	3,2	60	2 393
Belgie	5,4	3,3	61	1 880
Maďarsko	5,4	3,4	63	983
Španělsko	6,0	3,0	50	1 176
Finsko	6,0	3,0	50	989
Ukrajina	6,2	3,2	52	1 108
Slovensko	6,6	3,3	50	1 273
Rakousko	6,6	3,0	45	1 385
Itálie	7,4	2,6	35	1 951
Irsko	7,4	2,6	35	1 009
Řecko	7,9	2,6	33	1 135
Polsko	8,4	2,2	26	1 078
Rumunsko	8,6	2,2	26	1 124
Celkem	***6,0***	***3,2***	***53***	***38 661***

Zdroj: Data výzkumu EVS 1999.

Tab. 3.3 Jak je důležitý Bůh v životě člověka v různých evropských zemích
Pozn.: Tabulka je uspořádána vzestupně podle průměrné důležitosti Boha v životě.

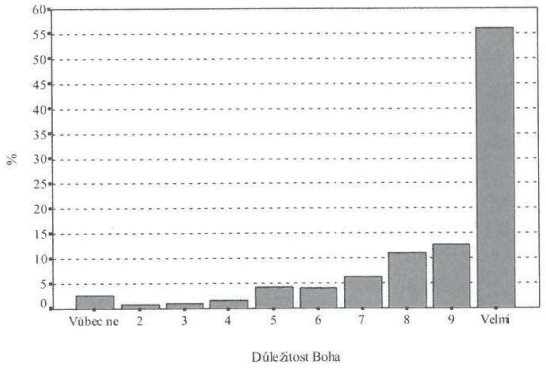
[95] Pozor ale, v analýze dat mějte neustále na paměti, že v sociologickém výzkumu pracujeme většinou s indikátory. I tato otázka je jen určitým indikátorem náboženské orientace, neboť ne všechna náboženství jsou založena na koncepci Boha, jak jej prezentuje křesťanství. Proto i ti, kdo říkají, že Bůh není v jejich životě vůbec důležitý, ještě nemusejí být ateisty. Kdo se chce o postojích k náboženství dozvědět více, nechť si přečte stať Lužného a Navrátilové v časopise *Sociální studia* 2001.

Více informací už z této analýzy asi nevytěžíme, což opět potvrzuje naše předchozí tvrzení, že třídění prvního stupně většinou nikdy žádné převratné poznatky nepřináší. Je to totiž především nástroj deskripce, ne skutečné analýzy.

Jiná situace ale nastane, pokud stejnou otázku položíme v sociologickém šetření v mnoha zemích. Pak získáme následující výsledek (viz tab. 3.3).

Z tabulky vyplývá, že ČR je zemí, kde respondenti v roce 1999 přisuzovali Bohu tu nejméně důležitou roli v jejich životě (ale všimněte si variačního koeficientu, který je ze všech zemí nejvyšší), blízko k nám má ještě Dánsko a Švédsko. Naopak velkou roli v životě člověka hrál Bůh u obyvatel Řecka, Polska a Rumunska (a poměrně nízký variační koeficient naznačuje nízký rozptyl dat). Taková komparativní data už mají velkou analytickou hodnotu, což je ale dáno částečně tím, že se de facto nejedná o třídění prvního stupně, ale o třídění stupně druhého (víte proč?).[96] Také si všimněme, že velikost výběrových souborů se v každé zemi pohybuje přinejmenším kolem tisícovky. Je to totiž počet, který již dovoluje dobrou a poměrně detailní analytickou práci (lze analyzovat nejen celek, tj. všechny dospělé české respondenty, ale i podskupiny, například jednotlivé vzdělanostní skupiny či kraje).

Pro ilustraci si ještě uveďme rozložení četností u otázky na důvěru v Boha, které získali naši rumunští kolegové (viz výstup 3.15). Když jej srovnáme s českým rozložením (viz výstup 3.14), vidíme, že rumunská a česká distribuce jsou téměř zrcadlovým obrazem jedna druhé. Zajímavý výsledek, není-liž pravda?



Zdroj: Data výzkumu EVS 1999.

Výstup 3.15 Důležitost Boha v životě jedince v Rumunsku v r. 1999

[96] Protože do tabulky jsme umístili dvě proměnné: „evropské země" a „důležitost Boha". O třídění druhého stupně a kontingenčních tabulkách pojednáme v kapitole 8.

Literatura

Hendl, J. (2004). *Přehled statistických metod zpracování dat*. Praha: Portál.

Lužný, D., & Navrátilová, J. (2001). Náboženství a sekularizace v České republice. *Sociální studia*, (6), 111–125.

Řehák, J. (1976). Základní deskriptivní míry pro rozložení ordinálních dat. *Sociologický časopis*, *12*(4), 416–431.

Swoboda, H. (1977). *Moderní statistika*. Praha: Svoboda.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading: Addison-Wesley.

Kapitola 4

Normální a standardizované normální rozdělení

4.1 Normální rozdělení

V předchozích lekcích jsme si ukázali, že předtím než začneme analyzovat data, je u proměnných měřených na intervalové úrovni (tedy proměnných spojitých, kardinálních) vždy dobré zjistit, jaký tvar má rozdělení jednotlivých znaků. Zajímá nás především, zdali má distribuce četností tvar rozdělení normálního. Ve statistice totiž provádíme řadu nejrůznějších testů – zde nemáme na mysli, že studenti jsou znovu a znovu zkoušeni z toho, co už ve statistice umí –, což znamená, že sledujeme, do jaké míry naše data odpovídají nějakému statistickému modelu (blíže k tomu v následující kapitole). Abychom mohli tyto testy provádět, musejí být splněny některé předpoklady – a právě předpoklad normálního rozdělení je často jedním z nich.

Normální rozdělení má podobu zvonovité křivky (však mu také angličtina říká *bell curve*, podobně němčina *Glockenkurve*) symetrické kolem střední osy (viz graf 4.1). Ve vědeckém jazyce se hovoří o **Gaussově křivce** (podle německého matematika a fyzika Karla Friedricha Gausse 1777–1855) nebo také o křivce normálního rozdělení.

Normální rozdělení je typické pro řadu biologických nebo psychických jevů, ale také pro některé vlastnosti sociální.[97] Podle francouzského matematika, statistika a astronoma belgického původu Adolpha Quételeta (1796–1874) normální rozdělení neznamená nic jiného než to, že příroda se snaží vytvořit ideální typ (reprezentovaný průměrem), avšak v různé míře (náhodně) chybuje.[98]

[97] Výraz „normální" je zde poněkud zavádějící – zvláště v sociálních vědách, kde mnoho proměnných je rozloženo jiným způsobem, takže mají podobu rozdělení ne-normálního. Slovo „normální" se v sousloví „normální rozdělení" vztahuje k staršímu významu „řídící se zákonem, předpisem nebo modelem".

[98] Reisenauer (1970) uvádí, že normální rozdělení je pozorováno při opakovaném měření téže veličiny za stejných podmínek. Jednotlivé naměřené hodnoty se v důsledku působení náhodných vlivů více či méně odchylují od skutečné hodnoty, jinými slovy jsou zatíženy tzv. náhodnými chybami.