LEKCE 02a
PŘÍKLAD

ROZLOŽENÍ VARIANT KATEGORIZOVANÉ PROMĚNNÉ (*FREQUENCIES*) A ČIŠTĚNÍ DAT

I. Čištění dat

Prvním krokem, který musíme udělat před jakoukoliv analýzou dat, je tzv. čistění dat. Nejedná se o nic jiného než o kontrolu dat – to je zdali při jejich nahrávání nedošlo k chybě, zdali jsme nenahráli jiné hodnoty, než které jsme zjistili ve výzkumu. Navíc –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 *outliers*, neboli extrémně odlišné hodnoty, čili „úleťáci"). Outliers většinou vznikají chybou při nahrávání: např. stiskem špatné klávesy, kdy např. při nahrávání hodnot pro pohlaví, které mohou být 1 nebo 2, omylem uhodíme na jinou klávesu a nahrajeme hodnotu 6 nebo přidáním řádu, když např. při nahrávání měsíčního příjmu respondenta, který je 15800 ve skutečnosti nahrajeme 158000 apod.

Čištění dat není příliš záživná činnost, 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 okřídlený počítačnický akronym GIGO znamenající *Garbage In, Garbage Out* (smetí dovnitř, smetí ven) a říkající, že pokud nahrajete špatná data, budou i vaše výsledky nutně špatné. Čištění dat probíhá ve dvou krocích:

1. Kontrola chybných dat
2. Nalezení chyby a její oprava

1. krok: Kontrola chybných dat

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 distribuce, rozložení nahraných hodnot se 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 proměnných nominálních a ordinálních a také u proměnných intervalových s malým počtem variant (např. počet dětí respondenta) je způsob kontroly odlišný od proměnných intervalových s velkým počtem variant (např. věk, příjem, IQ skóre, skóre v přijímacím testu atd.). Nazývejme pro zjednodušení tu první skupinu dat daty kategorizovanými, tu druhou pak daty nekategorizovanými.

A) Kontrola kategorizovaných dat
Data kontrolujeme tím způsobem, že si necháme udělat rozložení četností jednotlivých proměnných. K tomu použijeme proceduru

Analyze – Descriptive Statistics – Frequencies

a v rámci Frequencies si ještě necháme vytisknout minimální a maximální hodnotu znaku.

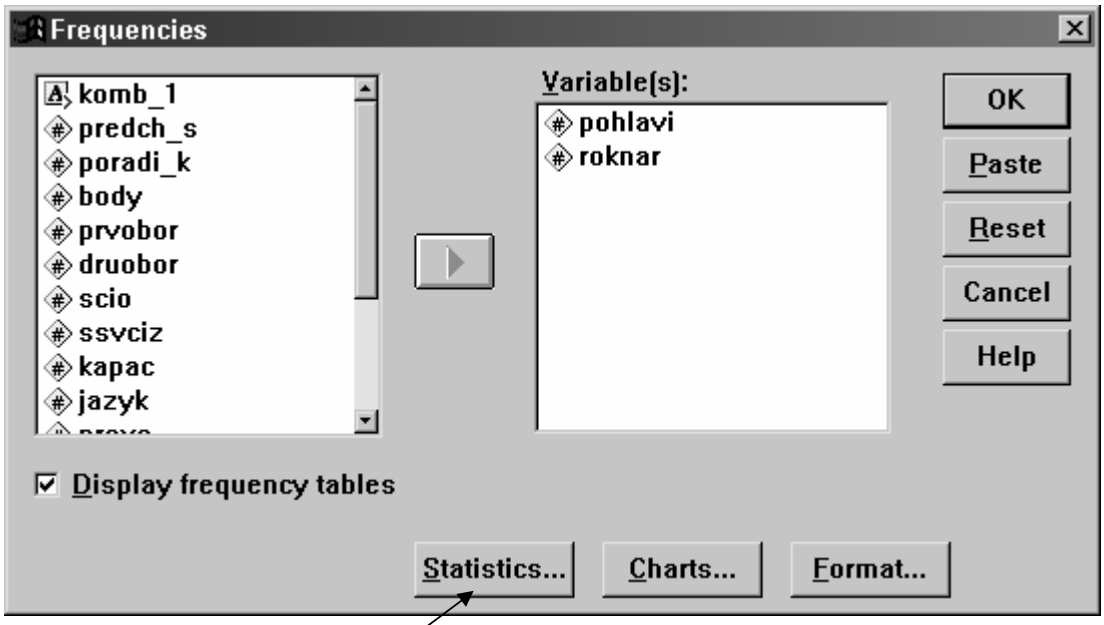
Ukažme si vše na příkladu. V našem souboru dat o přijímacím řízení (viz soubor *prij-error.sav*)[1] zkontrolujeme proměnné pohlaví a rok narození. Rok narození bývá ve většině výzkumů proměnná, která má velký počet variant, takže bychom ji měli chápat jako proměnnou nekategorizovanou, ale v našem případě – jelikož se jedná o uchazeče o prezenční studium na VŠ – bude mít variant jenom omezený počet (dokážete říci proč?). Je to tedy vhodná proměnná pro tuto proceduru.

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

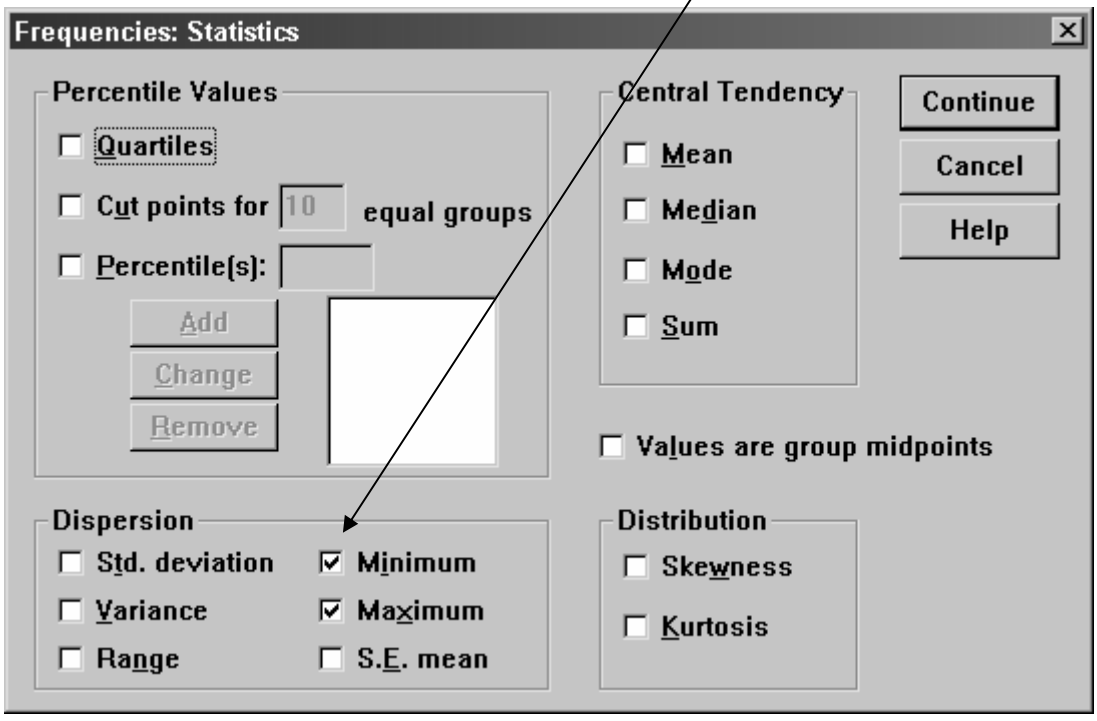
Nuže, jak postupujeme?

Analyze – Descriptive Statistics – Frequencies

Ve **Frequencies** klikneme na jména těch proměnných, které chceme kontrolovat, a přesuneme je do okna **Variable(s)**. V našem příkladě to jsou *pohlavi* a *roknar*.



Klikneme dále na tlačítko **Statistics** – a zde si zvolíme nalezení minimální a maximální hodnoty.



Po kliknutí na tlačítko **Continue** a pak na **OK** získáme následující výstup:

Výstup P2_1:

a)

Statistics

		POHLAVI	ROKNAR Rok narození
N	Valid	180	180
	Missing	0	0
Minimum		0	1879
Maximum		3	1991

b)

POHLAVI

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	3	1,7	1,7	1,7
	1 muz	92	51,1	51,1	52,8
	2 zena	82	45,6	45,6	98,3
	3	3	1,7	1,7	100,0
	Total	**180**	**100,0**	**100,0**	

c)

DAT_NARO Datum 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 a) je důležitý. Vidíme v něm především, že u obou proměnných je počet případů 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), je to samo o sobě důležitá informace, že něco není s příslušnými proměnnými v pořádku a je třeba zjistit, proč tam ty chybějící hodnoty jsou. Dále vidíme, že u proměnné pohlaví je minimální hodnota 0 a maximální 3, což jsou zřetelně omyly, neboť interval, v němž se hodnoty této proměnné mohou pohybovat je <1;2>. U proměnné rok narození jsou rovněž chyby. Minimální hodnota je rok 1879 (tedy 119letý uchazeč o studium) a maximální hodnota je 1991 (tedy 8letý uchazeč).

Ve výstupu b) máme rozložení proměnné pohlaví. Vidíme, že pohlaví s hodnotou 0 mají tři případy a s hodnotou 3 rovněž tři. Ve výstupu c) je rozložení hodnot roku narození. Jeden uchazeč má rok narození 1879, o němž už víme, že to je očividně chybný údaj, jeden 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. Další dva případy s rokem narození 1990 a 1991 jsou ale zcela jistě omyly.

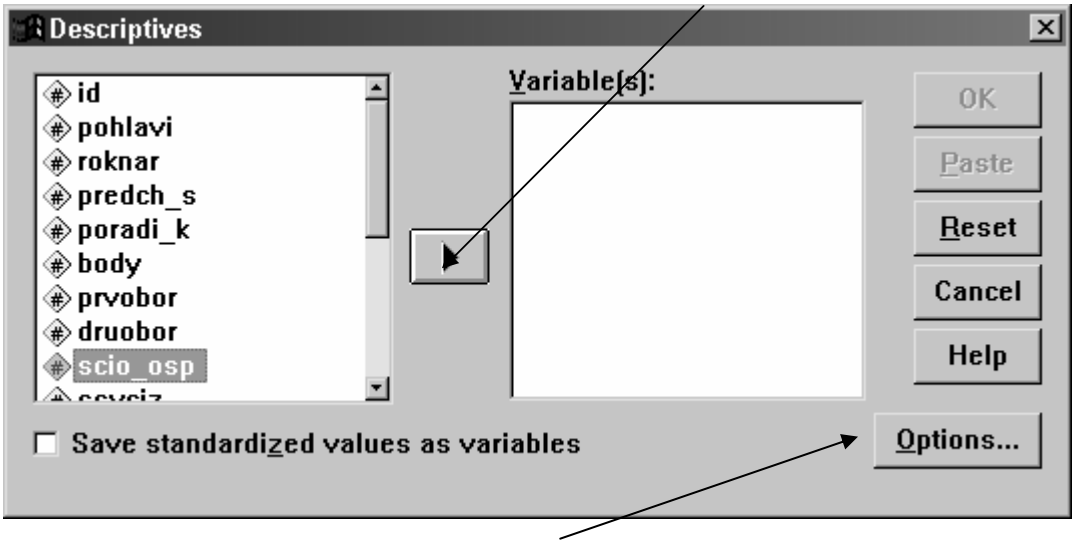
Kontrola nekategorizovaných dat

Nekategorizovaná data s velkým rozsahem hodnot (s velkým množstvím variant) nemá cenu kontrolovat prostřednictvím procedury **Frequencies** – dostali bychom totiž příliš mnoho řádků. Namísto Frequencies proto použijeme procedury

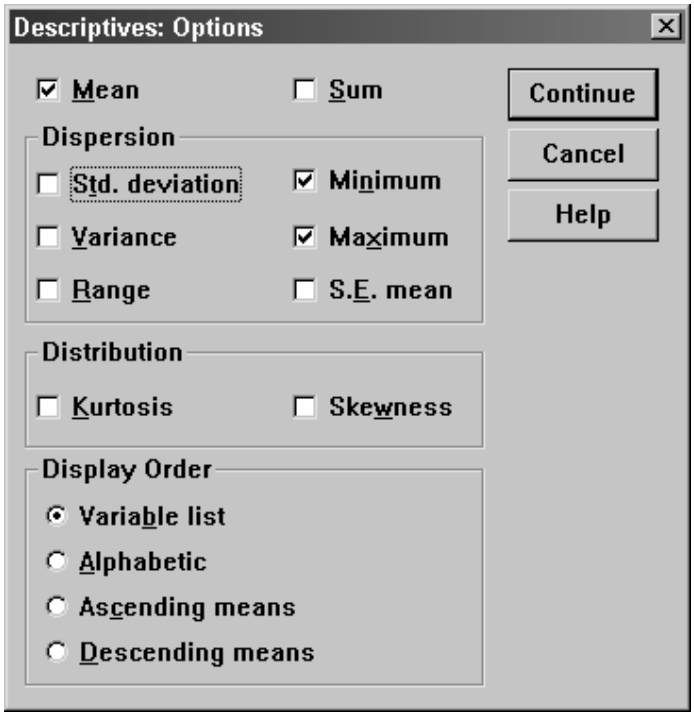
Analyze – Descriptive Statistics – Descriptives

Opět příklad. V našem datovém souboru z výsledků přijímacího řízení 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 ji, 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** a v dialogovém okně si zaklikneme požadavek na minimální a maximální hodnotu a také na průměr **(Mean).**



Poté klikneme na **Continue** a pak na **OK**. Ve výstupu se nám objeví následující tabulka (viz výstup P2_2.

Výstup P2_2:

Descriptive Statistics

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

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. 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 skončili první krok čištění dat a musíme postoupit ke kroku druhému.

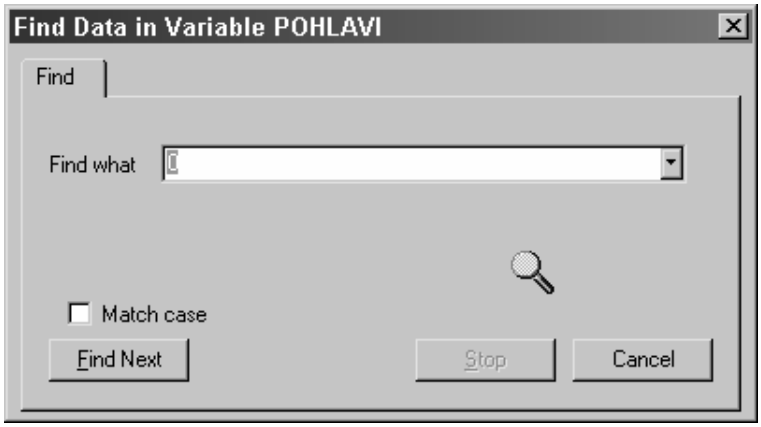
2. krok: Nalezení chyb a jejich oprava

Nyní tedy víme, že v našem datovém souboru jsou chyby, které je třeba opravit. Máme dvě možnosti. Pokud máme dostatečně velký soubor (např. 2 400 respondentů), můžeme si klidně dovolit těchto několik chybných případů obětovat a chybné hodnoty prohlásit (to je rekódovat) jako hodnoty chybějící (*missing values*) – jak to udělat si ukážeme v kapitole 4. *Missing values* pak nevstupují do žádných analýz.

Máme-li relativně malý soubor (do tří čtyř stovek), měli bychom chyby opravit podle skutečných hodnot.[2]

Vyhledat chybu není příliš obtížné. Hledáme ji přímo v datech, v datovém editoru **(Data View)**. Postupujeme následovně:

1. 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 vysvítil černě. Pak klikneme na **Edit** a na **Find**. Do příslušného okénka vepíšeme chybnou hodnotu, kterou chceme nalézt. My budeme nejdříve hledat 0.



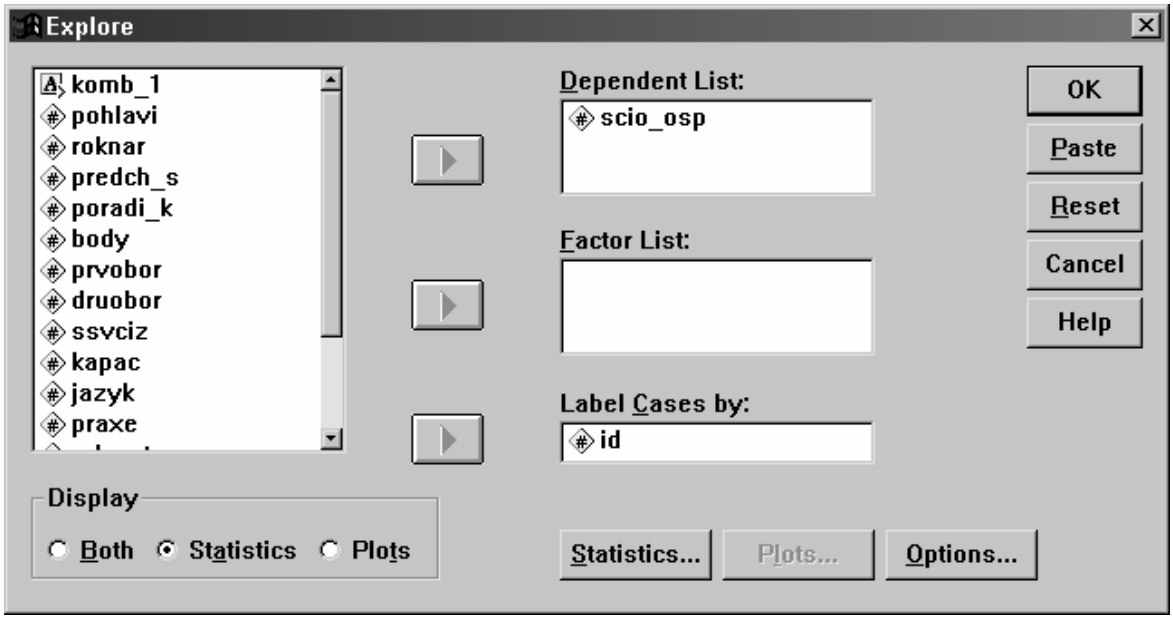
[2] Osobně doporučuji, abychom chyby opravovali i ve velkých souborech. Sběr dat je velmi nákladný a každý údaj, který je nevyužit, je plýtváním.

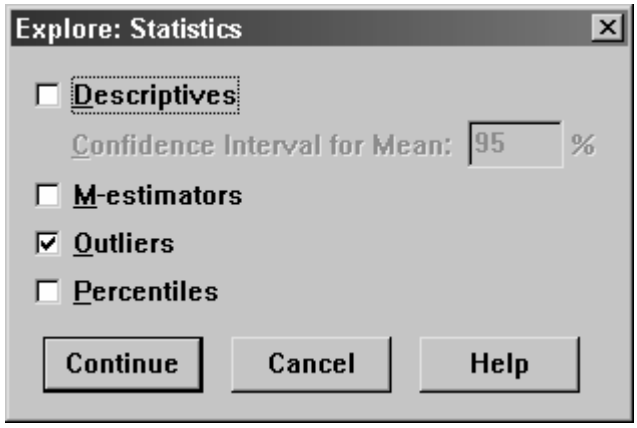
Klikneme myší na **Find Next** – ve vyčerněném datovém sloupci se objeví bílá buňka s hodnotou 0. Podíváme se do sloupce ID (identifikace), abychom zjistili, o který případ se jedná. V našem případě je to uchazeč č. 15. Není nyní nic lehčího než jít do testů z přijímacího řízení, vyhledat uchazeče s číslem 15 (jistě máme všechny testy dobře archivovány a seřazeny podle čísla identifikace), zjistit, jakého je pohlaví a do vysvíceného políčka vepsat správnou hodnotu. Jelikož víme, že v datech byly hodnoty 0 celkem třikrát , klikneme opět na **Find Next**, zjistíme identifikační číslo uchazeče a 0 opravíme. Totéž pak uděláme ještě potřetí. Stejným způsobem opravíme i chybné hodnoty 3.

Máme-li nekategorizovanou 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 (pokud tam jsou), použijeme proceduru **Explore**. Postupujeme takto:

Analyze – Descriptive Statistics – Explore.

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





Po klinutí na **Continue a OK** získáme výstup 52_3:

Výstup 52_3:

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

V něm jsou důležité poslední dva sloupečky nadepsané *Value* a *ID*. Sloupec Value udává pět nejvyšších hodnot proměnné (v horní polovině tabulky nad čarou, která je označena jako *Highest*), které se v souboru vyskytují 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)[3], hodnotu 645 uchazeč č. 150 a hodnotu 181 uchazeč č. 180. To jsou zřetelné překlepy. Hodnota 93 uchazeče č. 5 je již v pořádku, neboť maximálním počet bodů v tesu byl 100. U nejnižších hodnot je hodnota 7 podezřelá a měli bychom ji zkontrolovat. Hodnota 52 je již očividně v pořádku.

V proměnné *scio_osp* jsme tedy detektovali celkem čtyři chyby, které musíme opravit způsobem popsaným v předchozím oddíle – pouze s tím rozdílem, že už nemusíme vyhledávat jejich identifikace v datové matici. To za nás udělala procedura **Explore** a **Outliers**.

II. Analýza dat – třídění prvního stupně (Frequencies)

Až poté, kdy jsme zkontrolovali všechny proměnné v souboru a data vyčistili, můžeme přistoupit k vlastní analýze. Začínáme vždy tzv. univariační analýzou, tedy tříděním podle jedné proměnné, tříděním prvního stupně.

Příklad P2.1: V mezinárodním komparativním výzkumu *European Values Study*, který v České republice provedl v roce 1999 Jan Řehák a Ladislav Rabušic (data sbírala agenrura SC&C) na reprezentativním souboru české dospělé populace (ve věku 18 let a starší) 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.

Třídění prvního stupně nominálních a ordinálních a intervalových proměnných s malým počtem variant získáme prostřednictvím procedury:

[3] 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 respondneta či případu (ID) je zbytečné. Ne, není to zbytečné a každá datová matice SPSS by jako první proměnnou měla mít právě ID. Pokud identifikační proměnná chybí, je nutné ji vytvořit.

Analyze – Descriptive Statistics – Frequencies

V našem případě jsme třídili proměnnou q46_3 a získali jsme tuto tabulku:

Tab. P2_4: Výstup z procedury Frequencies, proměnná Q46_3.

Q46_3 Většina žen touží po domově a dětech

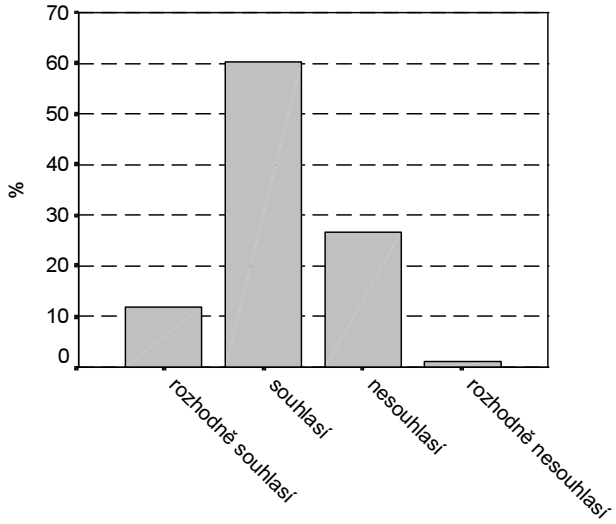
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 rozhodně souhlasí	213	11,2	12,0	12,0
	2 souhlasí	1070	56,1	60,1	72,1
	3 nesouhlasí	474	24,9	26,6	98,7
	4 rozhodně nesouhlasí	22	1,2	1,3	100,0
	Total	1780	93,3	100,0	
Missing	-2 neodpověděl/a	8	,4		
	-1 neví	120	6,3		
	Total	128	6,7		
Total		1908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

Tabulka říká, že na tuto otázku z celkového počtu 1908 dotázaných odpovědělo 93,3 % respondentů (což bylo 1780 osob) a 6,7 % respondentů (to je 128) se nerozhodlo ani pro jednu z nabízených variant (jejich odpovědi chybějí, proto jsou v tabulce umístěny do oddílu *Missing values*). Jelikož z výzkumného hlediska mají pro nás většinou význam pouze odpovědi těch, kdo mají na položenou otázku nějaký názor, pracujeme obvykle s údaji, které jsou umístěny ve sloupci *valid percent* (platná procenta). Vidíme, že 12 % respondentů *rozhodně souhlasí* s tím, že ženy především touží po domově a dětech. 60 % pak s tímto názorem *souhlasí*.[4] Celkem si tedy 72 % (12 + 60,1) českých respondentů myslí, 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). Opačný názor má 28 % respondentů (26,6 + 1,3).

Pokud vás při čtení těchto výsledků napadlo, že by bylo asi zajímavější, než znát rozložení tohoto postoje pro všechny respondenty, kdybychom měli tabulku, jak se k tomuto výroku staví ženy a jak muži nebo věkově mladší respondenti ve srovnání s těmi věkově staršími, pak jste na správné analytické stopě. Skutečně, třídění prvního stupně neboli **Frequencies** nemají v sociologických analýzách příliš velký věcný význam. Jsou ovšem neocenitelným pomocníkem při kontrole dat a také dobře slouží jako základní informace před složitějšími analýzami. A samozřejmě, jsou hlavním typem výstupů ve výzkumech veřejného mínění, kde nám např. říkají: „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“. Mnozí analytikové zastávají názor, že je mnohem lepší prezentovat výsledek analýzy prostřednictvím obrázku než v číslech. Předchozí tabulku tedy uvádíme v grafické podobě.

Obr. P2_1: Ukázka grafického výstupu (bar chart)

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



SPSS umí spoustu grafů (viz modul **Graphs**), ale bohužel nejsou vůbec hezké. Proto doporučujeme výsledná data vepsat do Excelu (tabulku z frequencies lze pomocí Ctrl–C a Ctrl-V zkopírovat a vložit přímo do Excelu, takže není nutné hodnotu přepisovat) a graf vytvořit jeho prostřednictvím. Grafy v SPSS lze editovat tak (dvakrát klikneme na obrázek), aby výstup splňoval českou normu pro prezentaci grafů. Jak se to dělá je předmětem příslušných cvičení ve výuce.

LEKCE 02b
PŘÍKLAD

V případě, kdy sledovaná proměnná je proměnnou ordinální s mnoha variantami nebo když se jedná o proměnnou intervalovou, třídění prostřednictvím *Frequencies* nemá smysl (dokážete odpovědět, proč tomu tak je?). U takových proměnných použijeme pro analýzu střední hodnoty a míry variability. Nejvíce se ovšem tento postup hodí pro kardinální proměnné. K výpočtu jsou k dispozici tři procedury: *Frequencies* (u nichž zaškrtneme políčko, že ve výstupu nechceme tabulku třídění, avšak že požadujeme výpočet statistik, popř. i grafické zobrazení), *Descriptives* a *Explore*. Ukažme si je postupně všechny.

Příklad P2.2: Na základě přijímacích zkoušek bylo na fakultu X přijato do bakalářského prezenčního studia přijato celkem 180 studentů. Proveďme analýzu jejich bodového zisku.

A) Výpočet prostřednictvím procedury *Frequencies*:

Tab. 2.2: Ukázka výstupu procedury *Frequencies*, *statistics*

Statistics

TESTYALL

N	Valid	180
	Missing	0
Mean		139,71
Std. Error of Mean		,59
Median		140,00
Mode		141[a]
Std. Deviation		7,86
Variance		61,83
Skewness		1,232
Std. Error of Skewness		,181
Kurtosis		6,215
Std. Error of Kurtosis		,360
Range		63
Minimum		120
Maximum		183
Percentiles	25	136,00
	50	140,00
	75	143,00

a. Multiple modes exist. The smallest value is shown

Co nám tato tabulka říká? Nejdříve se v datech musíme zorientovat. Teoreticky mohli uchazeči o studium získat v přijímacích písemných testech 0 – 200 bodů (tuto informaci nevyčtete z tabulky, to je danost přijímacího řízení fakulty X). Podívejme se do spodní části tabulky na údaje o dosaženém minimu a maximu. Vidíme, že minimální počet bodů, který ještě stačil k přijetí, byl 120, a že získaný maximální počet bodů byl 183. V tomto intervalu 120–183 se tedy pohyboval bodový zisk přijatých studentů .

Průměrné skóre (*mean*) mělo hodnotu 139,7 bodů, nejčastějším bodovým ziskem (*mode*) bylo 141 bodů. Údaje o percentilech říkají, že 25 % přijatých získalo mezi 120 –136 body (zde jsme si spojili informaci o minimální dosažené bodové hodnotě s údajem o 25 percentilu (také se mu říká první nebo dolní kvartil), dalších 25 % přijatých mělo bodový zisk mezi 136 a 140 body – hodnota 50. Percentilu je současně mediánem (*median*), který říká, že 50 % uchazečů získalo méně než 140 bodů a dalších 50 % uchazečů získalo více než 140 bodů. 75 % uchazečů pak získalo do 143 bodů. Nejlepší čtvrtina uchazečů pak měla bodový zisk mezi 143 body a 183 body.
Údaj o průměru by neměl být nikdy používán osamoceně bez toho, že bychom jej doplnili informací o variabilitě hodnot znaku. Základní mírou variability je rozptyl (*variance*), v našem případě má hodnotu 61,8. Pro analytické účely není příliš informativní, mnohem lepší je používat jeho druhou odmocninu, směrodatnou odchylku (*std. deviation*). Ta je 7,86.[5] Naznačuje tedy, že bodový zisk jednotlivých uchazečů byl poměrně vyrovnaný a že rozptyl v datech nebyl příliš velký. Čím je hodnota směrodatné odchylky nižší, tím jsou data více homogenní – hodnota průměru je v takovém případě údajem, který dobře charakterizuje data. Dobrým indikátorem toho, jak jsou data rozptýlena, je srovnání průměru, mediánu a modu – v našem případě jsou si všechny tři údaje velmi podobné, takže data jsou vskutku poměrně homogenní. Pokud by byla směrodatná odchylka velká, hodnoty průměru, modu a mediánu by se odlišovaly. Znamenalo by to např., že někde v datech je několik atypických případů (*outliers* – „úleťáků“ s odlehlými hodnotami). V takovém případě není dobré používat průměr, neboť ten je těmito odlehlými hodnotami ovlivněn, přednost je třeba dát mediánu.

[5] 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.

Jiným indikátorem rozptylu v datech je ***variační koeficient***, což je jedna z nejlepších měr relativní variability. Je to poměr směrodatné odchylky k aritmetickému průměru, násobený 100 (je nutné ho vypočítat na kalkulačce, SPSS nemá tento výstup zabudován). Náš variační koeficient je (7,86/ 139,7)*100 = 5,6 %. Variační koeficient je výborným nástrojem při srovnání dvou souborů. Představme si jiný soubor, např. přijaté studenty na Fakultu sociálních věd UK, kteří by dělali stejné přijímací testy jako uchazeči o studium na FSS. 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. Variční koeficient pražských přijatých by tedy byl 14,0 %, tedy více než dvojnásobný. V Praze byl tedy výkon v testech našich fiktivních přijatých mnohem heterogennější a možná, že hodnota průměru byla ovlivněna několika málo studenty, kteří získali vysoký počet bodů, zatímco zbytek mohl mít horší výkon než v Brně. K tomu abychom tuto otázku vyřešili bychom museli srovnat údaje o mediánu a o percentilech anebo si udělat některé grafické analýzy (viz oddíl C níže).

Variační koeficient je také dobré použít např. při srovnávání rozptýlenosti hodnot proměnných měřených v nestejných jednotkách. Např. budeme-li srovnávat homogenitu americké a české populace z hlediska příjmů, budeme mít české příjmy v korunách a měsíční, zatímco americké budou v dolarech a za rok.[6]

B) Výpočet prostřednictvím procedury *Descriptives*:

Tab. 2.3: Ukázka výstupu procedury *Descriptives*

Descriptive Statistics

		TESTYALL	Valid N (listwise)
N	Statistic	180	180
Range	Statistic	63	
Minimum	Statistic	120	
Maximum	Statistic	183	
Mean	Statistic	139,71	
	Std. Error	,59	
Std.	Statistic	7,86	
Variance	Statistic	61,827	
Skewness	Statistic	1,232	
	Std. Error	,181	
Kurtosis	Statistic	6,215	
	Std. Error	,360	

Výstup z procedury *Descriptives*[1] přináší v podstatě stejnou druh informací jako procedura *Frequencies*. Není zde ale např. možnost volit si výpočet percentilů, naopak ale umí přetvořit hodnoty proměnné do tzv. Z skórů (blíže o nich v lekci o standardizovaném normálním rozložení).

C) Výpočet prostřednictvím procedury *Explore*:

Tab. 2.4: Ukázka výstupu procedury *Explore*

a)

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
TESTYALL	180	100,0%	0	,0%	180	100,0%

b)

[6] Je ovšem pravda, že v tomto případě bychom asi použili Giniho koeficientu, který je pro zachycení příjmového rozložení ve studiích o příjmové nerovnosti velkmi často užíván.

Descriptives

			Statistic	Std. Error
TESTYALL	Mean		139,71	,59
	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		7,86	
	Minimum		120	
	Maximum		183	
	Range		63	
	Interquartile Range		7,00	
	Skewness		1,232	,181
	Kurtosis		6,215	,360

c)

Percentiles

	Weighted Average(Definition 1)	Tukey's Hinges
Percentiles	TESTYALL	TESTYALL
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	

d)

Extreme Values

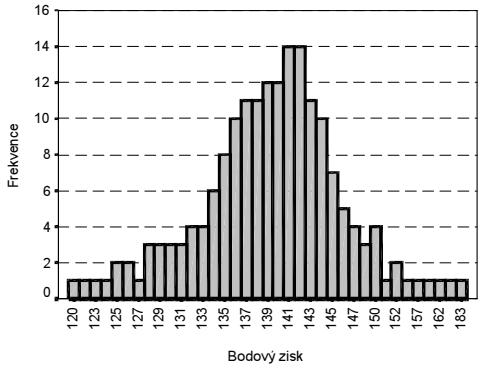
			Case Number	Value
TESTYALL	Highest	1	5	183
		2	64	172
		3	180	162
		4	173	161
		5	174	157
	Lowest	1	98	120
		2	97	121
		3	103	123
		4	16	124
		5	1	,[a]

a. Only a partial list of cases with the value 125 are shown in the table.

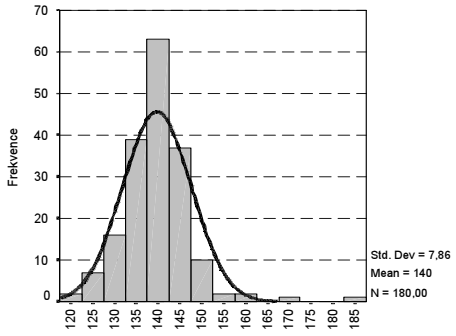
Výstup z této procedury má několik částí a ty obsahují některé nové informace. V tabulce b) je to např. údaj o intervalu spolehlivosti průměru (*95 % confidence interval for mean*), dále údaj o hodnotě průměru, pokud bychom soubor ořezali o 5 % nejnižších hodnot a 5 % nejvyšších hodnot (je tedy počítán z 90 % dat, která leží uprostřed tohoto intervalu). V našich datech není v podstatě rozdílu mezi „standardním“ průměrem a průměrem „ořezaným“, což je další důkaz toho, že v datech není příliš mnoho extrémních hodnot. V tabulce je i hodnota interkvartilového rozpětí (*interquartile range*), což je rozdíl mezi hodnotou dolního a horního kvartilu.

Tabulka c) uvádí hodnoty některých percentilů a tabulka d) případy pěti nejnižších a pěti nejvyšších hodnot. Z ní např. můžeme zjistit, že vůbec nejvyššího bodového zisku u přijímacích zkoušek získal uchazeč(ka) č. 5 a naopak nejnižší bodový zisk uchazeč(ka) č. 98. Jak jsme pravili již dříve, vždy je dobré, pokud to jde, samozřejmě, doplnit analýzu ještě o grafické výstupy. Procedura *Frequencies* umí vyrobit jednak sloupkový graf (viz obr. 2.2), ale umí také histogram s proloženou křivkou normálního rozložení (viz obr. 2.3).

Obr. 2.2: Ukázka grafického výstupu (bar chart) procedury *Frequencies*



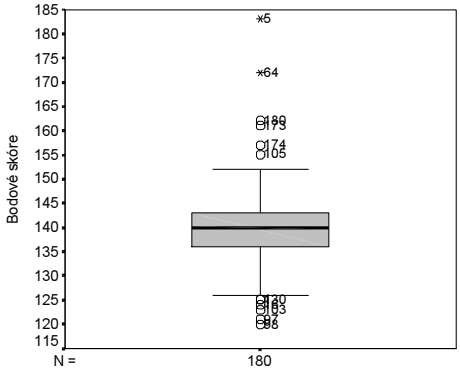
Obr. 2.3: Ukázka grafického výstupu (*histogram*) procedury *Frequencies* – histogram s proloženou křivkou normálního rozložení



Oba obrázky naznačují, že rozložení je přibližně normální (existuje test, který normalitu potvrdí či odmítne matematicky, ale o tom více v lekci 5). Není se čemu divit, vždyť intelektové schopnosti i těch, kdo jsou přijati, jsou rozloženy normálně, což znamená, že i mezi přijatými ke studiu na VŠ jsou nadaní a nadanější.

Procedura Explore umí ještě jednu velmi dobrou grafickou analýzu. Ukazuje ji obr. 2.4.

Obr. 2.4: Ukázka grafického výstupu procedury *Explore – Box and Whiskers*



Obrázek 2.4 je velmi informativní. Ukazuje, že data jsou velmi těsně rozložena kolem mediánu (tučná čára uprostřed krabičky) a že interkvartilové rozpětí je úzké (vertikální délka krabičky, v našem případě je to 7 bodů). V krabičce leží 50 % všech případů. Dolní hrana krabičky je 25. percentil a horní hrana 75. percentil. Dolní „vousy“ (*whiskers*) mají hodnotu 1,5 násobku interkvartilového rozpětí mínus hodnotu dolního kvartilu. V našich datech tato hodnota činí 136 – (1,5 x 7) = **125,5**. Horní vousy naopak hodnotu 1,5 násobku interkvartilového rozpětí plus hodnotu horního kvartilu. To je 143 + (1,5 * 7) = **153,5.** V grafu vidíte, že v 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 3 násobkem dolního či horního kvartilu) jsou hodnotami odlehlými (*outliers*). V našem případě je 1,5 násobek 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 *o* s číslem případu - vidíme tedy, že např. student/ka č. 174 má vysokou odlehlou hodnotu, zatímco student/ka č. 98 má nízkou odlehlou hodnotu (tuto informaci uvádí také tabulka 2.4c).

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 jsou studenti č. 64 a 5, jejichž hodnoty jsou extrémně vysoké – samozřejmě relativně, to je vzhledem k ostatním výsledkům.

* * *

V sociologickém výzkumu nemáme k dispozici data intervalová příliš často, většinu sociálních vlastností totiž neumíme na intervalových škálách změřit. Proto často pracujeme s daty ordinálními, u nichž alespoň konstruujeme dlouhé stupnice měření, jak ukazuje příklad P2.3

Příklad P2.3: 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	

(Pozn. Toto je častý způsob měření některých znaků. Tím, že takto měřená ordinální proměnná má mnoho stupňů měření, mění se proměnnou semi-intervalovou, u níž již má smysl používat mnohé statistické operace, které jsou určeny pro intervalové proměnné).

Výpočet prostřednictvím *Frequencies*:

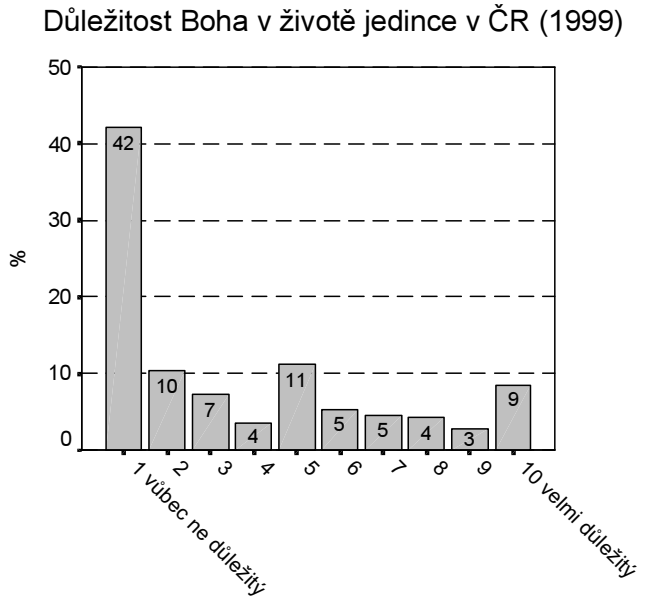
Tab. 2.5: Ukázka výstupu procedury *Frequencies*, *statistics*

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

Tabulku je vždy dobré doplnit ještě grafem, abychom si učinili představu, jak jsou data rozložena. Tvar rozložení je totiž důležitý, jak se dozvíte v pasážích o normálním rozložení. Obr. 2.4: Ukázka grafického výstupu procedury *Frequencies*, *charts*.



Jaké poznatky lze z těchto informací získat? Především vidíme, že procentuální rozložení odpovědí na tuto otázku je velmi nerovnoměrné (a má velmi daleko k rozložení normálnímu). Nejčastější odpovědí byla varianta „Bůh není v mém životě vůbec důležitý“ (42 % respondentů), proto má také modus (*mode*) v tabulce 2.2 hodnotu 1. Tato informace naznačuje, že značná část české populace není nábožensky založena.[7] Potvrzují to i další údaje: hodnota mediánu (*median*) 2 říká, že 50 % respondentů nemělo vyšší hodnotu tohoto znaku než 2 a průměrná hodnota všech respondentů je 3,6. 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 %).

Více informací už z této proměnné 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.

Pokud ale tuto otázku položíme v mnoha zemích a získáme následující výsledek (viz tab. 2.6), pak je to úplně jiná káva. Vidíme, že ČR je zemí, kde respondenti přisuzují Bohu tu nejméně důležitou roli v jejich životě (ale všimněte si variačního koeficientu), blízko k nám má ještě Dánsko a Švédsko. Naopak velkou roli v životě člověka hraje Bůh u obyvatel Řecka, Polska a Rumunska (a poměrně nízký variační koeficient naznačuje nízký rozptyl dat). Taková 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?). Také si všimněte, že pokud chcete získat pro nějakou populaci reprezentativní soubor, musí se velikost vzorku pohybovat minimálně kolem tisícovky respondentů.

Tab. 2.6: Jak je důležitý Bůh v životě člověka v různých evropských zemích

Země	Průměr	Směrod.	Variační	A

[7] 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 s Navrátilovou v časopise *Sociální studia* 2001.

		odchylka	koeficient	
ČR	3,6	3,1	86	1 846
Dánsko	4	2,8	70	1 001
Švédsko	4,1	3	73	995
Francie	4,4	3	68	1 580
Velká Británie	4,9	3,2	65	960
SRN	5	3,1	62	1 988
Slovinsko	5	3,2	64	980
Nizozemsko	5	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	3	50	1 176
Finsko	6	3	50	989
Ukrajina	6,2	3,2	52	1 108
Slovensko	6,6	3,3	50	1 273
Rakousko	6,6	3	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***

Pramen: EVS 1999

Obr. 2.5 Důležitost Boha v životě jedince v Rumunsku (1999)