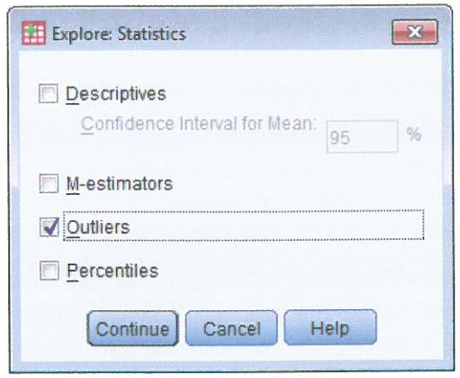


Obr. 3.8b Zadávání příkazu pro vyhledání problematických hodnot proměnné

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)
$norm.nomvar = K \times$ 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 × 0,722 / (5 − 1) = 3,61 / 4 = **0,903** (viz vzorec 3.3)

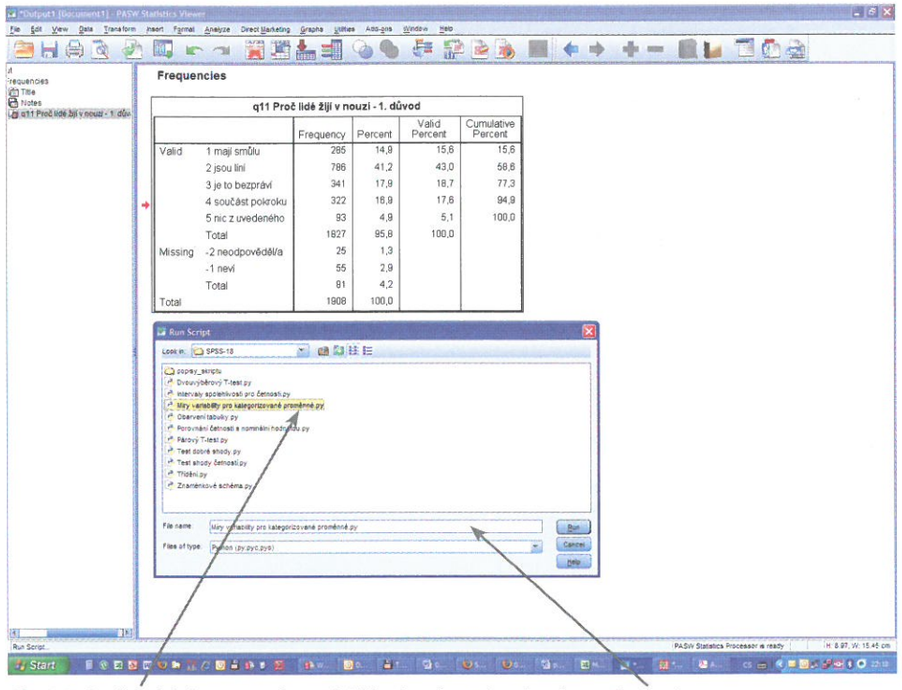
Z vypočtených údajů vyplývá, že míra koncentrace není vysoká, data jsou rozptýlena do všech kategorií.

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 programky, 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 programky 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.

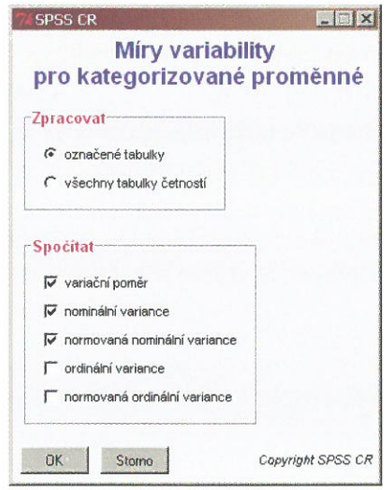
Skript máme i na míry variability nominální proměnné. Spustíme jej následovně: Necháme si udělat nám již známé rozložení četností proměnné *q11*: *Analyze – Descriptive Statistics – Frequencies*. Ve výstupu označíme tabulku kliknutím levým tlačítkem myši. U tabulky se objeví tučná červená šipka (viz 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 neodpov ěděl/a	5	,2		
	-1 nev í	14	,7		
	Total	19	1,0		
Total		1908	100,0		

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

$Dorvar = 2 \times 0{,}56 = 1{,}121$ (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**

$norm.dorvar = 2 \times 1{,}121 / (5 - 1) = 2{,}242 / 4 = \mathbf{0{,}561}$ (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*)

$$\bar{x} = \frac{\sum x}{n}. \tag{3.6}$$

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:

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

– **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.

$$s = \sqrt{s^2} \tag{3.8}$$

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

$$V = \frac{s}{\bar{x}} \times 100 \tag{3.9}$$

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	=	$\sqrt{70,2}$ = 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.

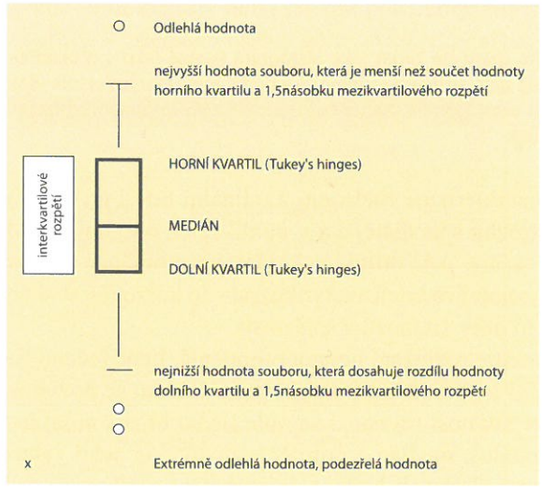
[90] Jen pro upřesnění, ale aby nás to nemátlo. Kvartily nejsou čtyři, ale tři; kvintilů není pět, ale jsou čtyři a decilů není deset, ale pouze devět. Je to jako při dělení úsečky na 4 stejně dlouhé úseky – stačí vám k tomu pouze 3 značky.

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):

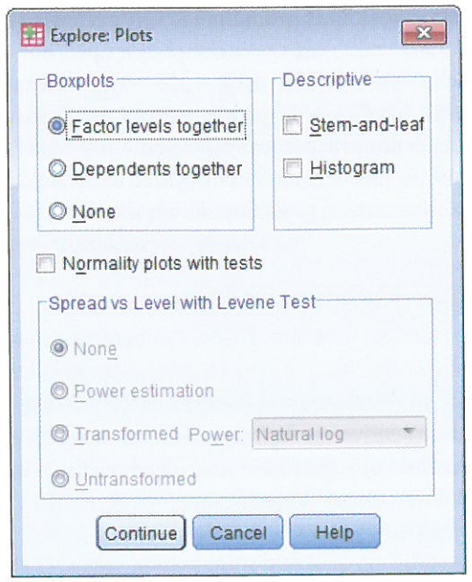


Obr. 3.12 Krabičkový graf

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

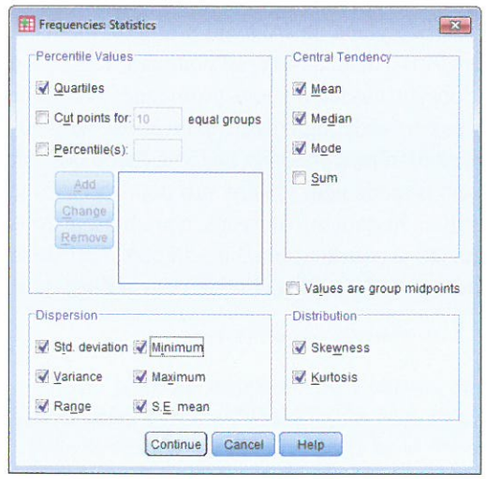
K výpočtům všech výše uvedených charakteristik centrální tendence a variability můžeme v SPSS využít tří procedur: *Frequencies*, *Descriptives* a *Explore*. Ukažme si je postupně všechny při aplikaci na jeden příklad.

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