LEKCE 02a
UNIVARIAČNÍ ANALÝZA KATEGORIZOVANÝCH DAT

Základní statistickou úlohou je popis stavu základního souboru

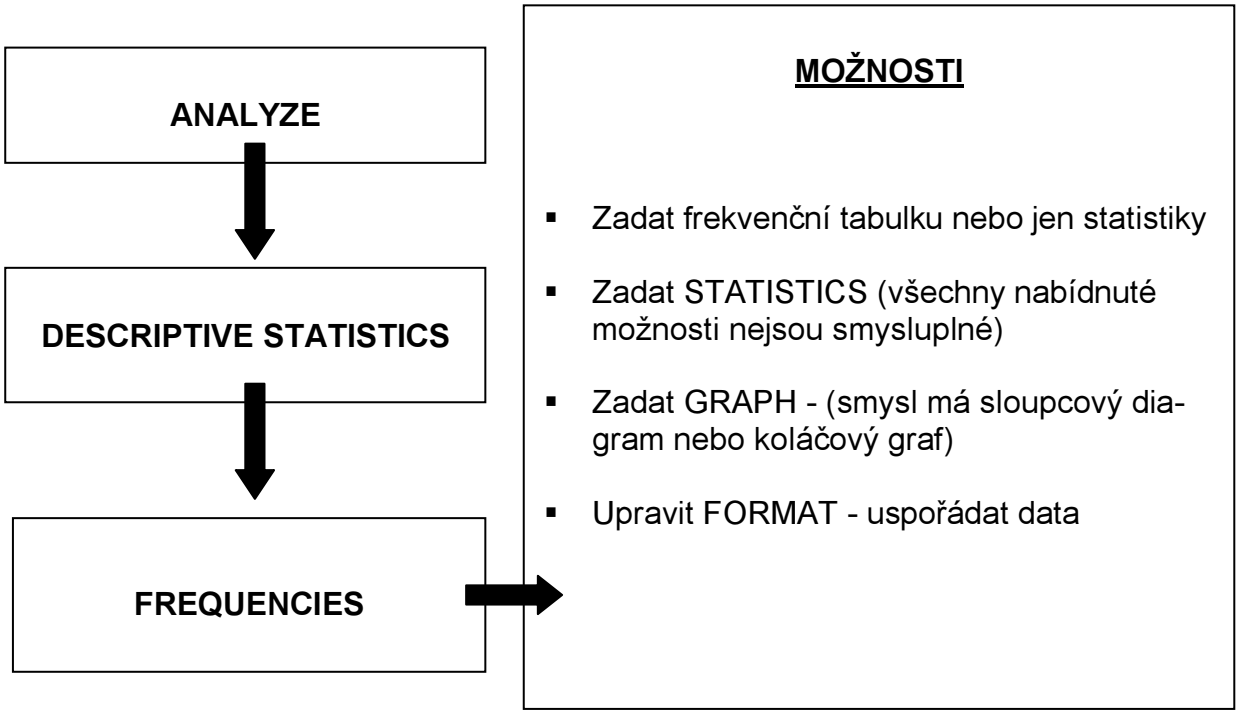
Východiskem je většinou výběrový soubor (odvozujeme popis základního souboru z popisu souboru výběrového). Statistický popis spočívá ve zjištění statistického rozložení Neboli rozdělení neboli distribuce četností u hodnot proměnné (znaku)

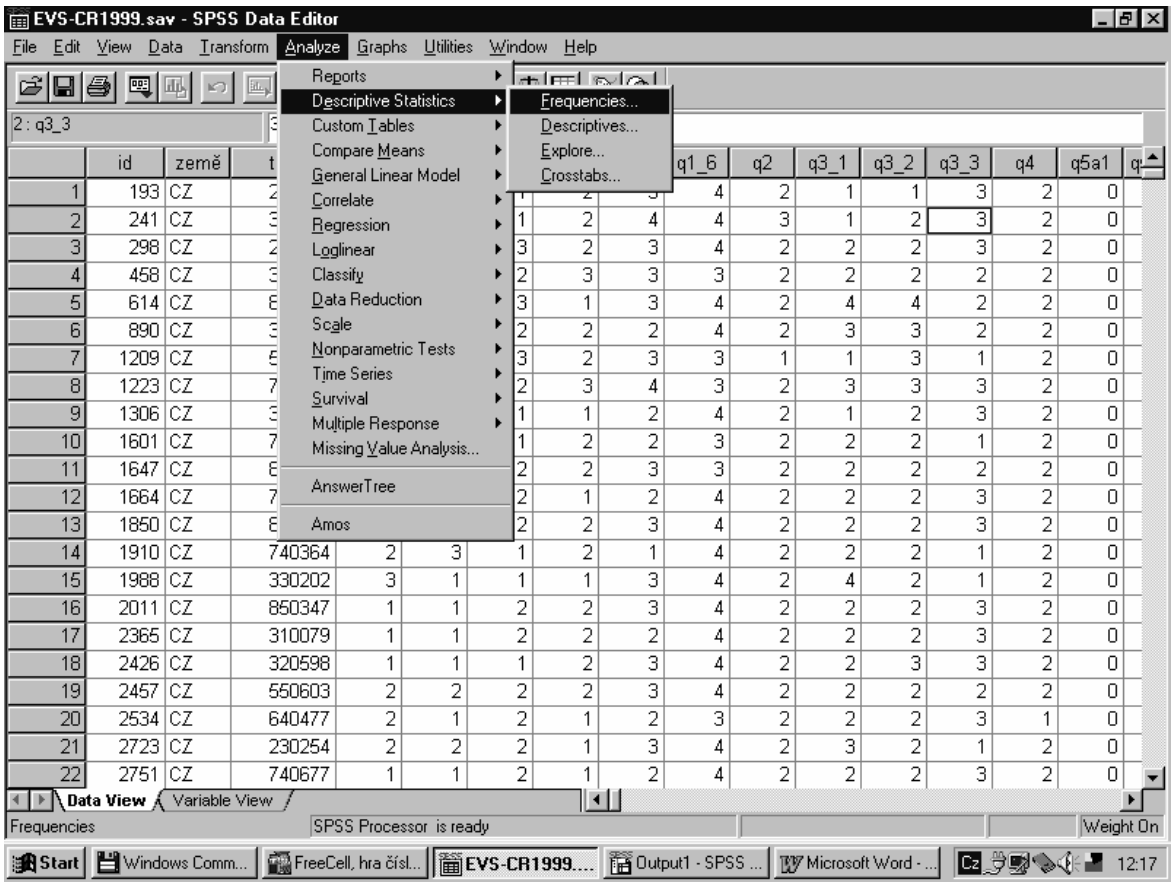
Statistické rozložení může být vyjádřeno v:

- Absolutních četnostech
  Kolik případů má danou vlastnost (z těch, jež jsou logicky v proměnné seskupeny)
  - Např. Kolik je v souboru mužů (žen).
  - Např. Kolik je v souboru osob s vysokoškolským vzděláním.

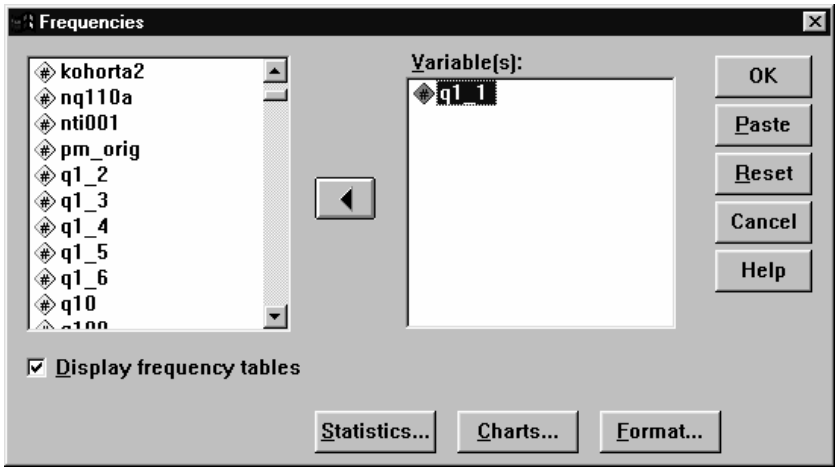
  Součet absolutních četností ve všech kategoriích (včetně chybějících hodnot) je velikostí (rozsahem) souboru.
- Relativních četnostech
  Jaký podíl představují případy mající danou vlastnost (z celku vlastností logicky v proměnné seskupených)
  - Např. Jaký podíl mužů (a jaký podíl žen) je v souboru.
  - Např. Jaký podíl osob s vysokoškolským vzděláním je v souboru.

  Součet relativních četností ve všech kategoriích dává 100%.
- Kumulativních relativních četnostech (nemají smysl u nominálních znaků)
  - Např. Jaký podíl osob alespoň s maturitou je v souboru.

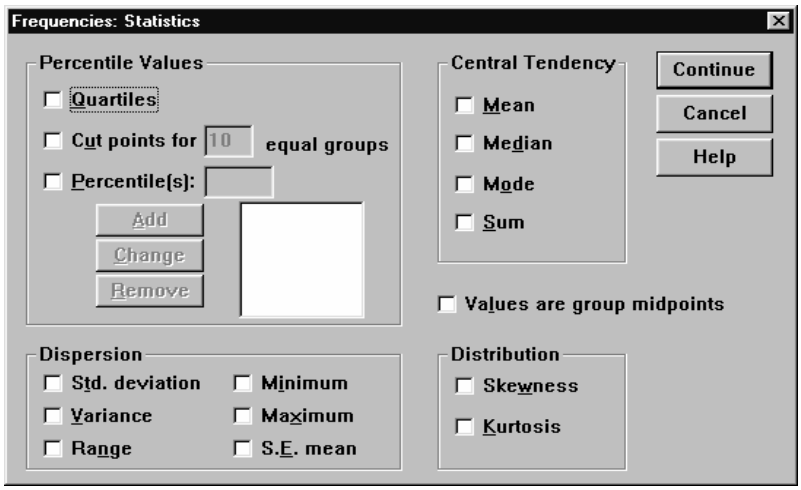




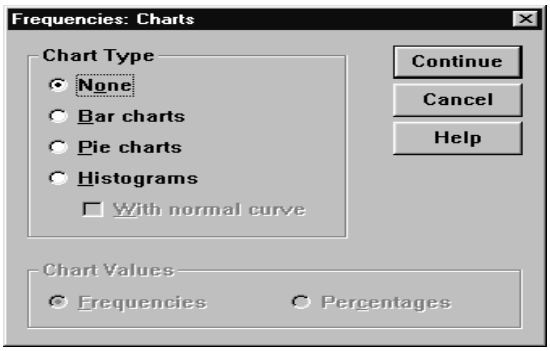
Zadání frekvenční tabulky zachycující rozložení dat (NOMINÁLNÍ A ORDINÁLNÍ PROMĚNNÉ).



Základním výstupem procedury FREQUENCIES je FREKVENČNÍ TABULKA: Můžeme ovšem zadat i další statistiky, které mají u daného typu proměnné smysl. U nominálních proměnných modus, u ordinálních modus a medián, pouze u kardinálních vedle modusu a mediánu i aritmetický průměr (a samozřejmě odpovídající míry rozptýlenosti) – ZDE OVŠEM NEMÁ SMYSL FREKVENČNÍ TABULKA A PROTO PRO FREKVENČNÍ ANALÝZU KARDINÁLNÍCH DAT POUŽIJEME RADĚJI PROCEDUR EXPLORE nebo DESCRIPTIVES, popřípadě potlačíme zobrazení frekvenční tabulky (odstranit zaškrtnutí *Display frequency tables*).



Zadat můžeme také jednoduchý graf zobrazující rozložení případů v kategoriích (absolutně či jejich procentuální podíl v celku).



Grafy ovšem často zadáváme raději v modulu menu *GRAPHS*. Je to jednoduché, chce to jen trochu experimentovat.

JAK ČÍST FREKVENČNÍ TABULKU

Příklad: Když zvážíte všechny okolnosti, řekl/a byste, že jste

Q4 Pocit štěstí celkově

		počet	podíl	validní podíl	kumulativní podíl
validní	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 ne šťastný/	26	1.4	1.4	100.0
	Total	1899	99.6	100.0	
chybějící hodnoty	-2 neodpověděl/a	5	.3		
	-1 neví	3	.2		
	celkem	9	.4		
celkem		1908	100.0		

základem pro výpočet jsou jen ti, kdo odpověděli

kódy vlastností

labels vlastností

podíl těch, kdo jsou alespoň „celkem šťastní“

VALIDNÍ RELATIVNÍ ČETNOSTI

Někdy nepočítáme podíl dané kategorie z celého souboru. Musíme si například poradit transformací proměnné tam, kde je její součástí kategorie, která nevstupuje do analýzy.

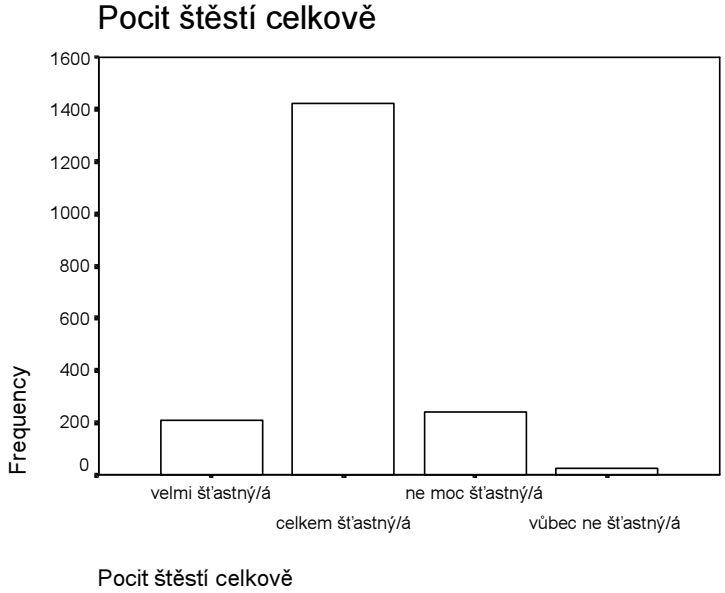
A6 JE NEKDO DOBROVOLNE CHUDY?

		Frequency	Percent	Valid Percent	Cumulative Percent	validní %
Valid	1 NIKDO	252	25,2	25,2	25,2	27,5
	2 VYJIMECNE	522	52,2	52,3	77,5	56,9
	3 NE MALO	143	14,3	14,3	91,8	15,6
	9 NEVIM	82	8,2	8,2	100,0	-
	Total	999	99,9	100,0		100,0
Missing	0	1	,1			
Total		1000	100,0			

Pokud by byla dále odstraněna varianta „ne vím“ - získáváme tak podíl postojů jen mezi těmi, kdo měli na věc názor.

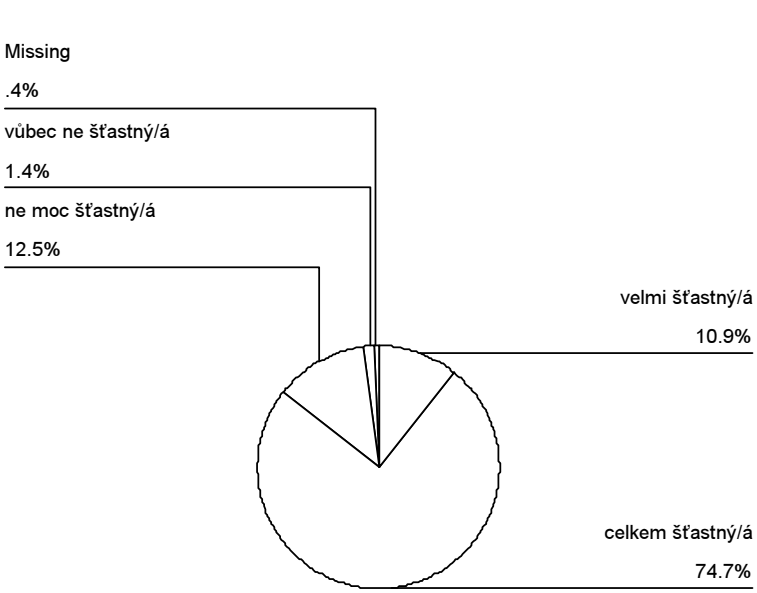
ZÁKLADNÍ ZOBRAZENÍ ROZLOŽENÍ ČETNOSTÍ KATEGORIZOVANÉ PROMĚNNÉ

SLOUPCOVÝ GRAF



Lze ho zadat v proceduře FREQUENCIES nebo v proceduře GRAPHS - BAR - SIMPLE

KOLÁČOVÝ GRAF



Lze ho zadat v proceduře FREQUENCIES nebo v proceduře GRAPHS - BAR - SIMPLE

CHARAKTERISTIKY ROZLOŽENÍ NOMINÁLNÍ PROMĚNNÉ

Střední hodnota:
- MODUS (nejčetněji obsazená kategorie neboli hodnota proměnné

Míra variability:
- Variační poměr = $1 - \frac{\text{četnost modální kategorie}}{\text{velikost souboru}}$

CHARAKTERISTIKY ROZLOŽENÍ ORDINÁLNÍ PROMĚNNÉ

Střední hodnota:
- MODUS
- MEDIÁN je číslo mediánové kategorie
  (MEDIÁNOVÁ KATEGORIE je ta, ve které je dosaženo 50% všech údajů, postupujeme-li od první kategorie výše)

Míra variability:
- VARIAČNÍ POMĚR
- Diskrétní ordinální variance (DORVAR)
- Normalizovaná diskrétní ordinální variance (NORM DORVAR)

POROVNÁVÁNÍ ROZLOŽENÍ
(tables - tables of frequencies)

MODUL TABLES dovoluje prezentovat v přehledné podobě frekvenční analýzu více proměnných

ANALYZE
TABLES
TABLES OF FREQUENCIES
Frequencies for: ANOMREC
In each tables: A75 (subjektivní třída)

Podíl osob s ruznou mírou anomie podle subjektivní třídy

	A75 TRIDA - SEBEZARAZENI SE							
	1 NIZSI TRIDA		2 NIZSI STREDNI		3 VYSSI STREDNI		4 VYSSI TRIDA	
	T01V0008 REKODOVANY INDEX ANOMIE		T01V0008 REKODOVANY INDEX ANOMIE		T01V0008 REKODOVANY INDEX ANOMIE		T01V0008 REKODOVANY INDEX ANOMIE	
	Count	%	Count	%	Count	%	Count	%
1 NIZKA	10	5,0%	107	18,0%	50	28,9%	7	46,7%
2 STREDNI	87	43,5%	260	43,8%	85	49,1%	6	40,0%
3 VYSOKA	103	51,5%	226	38,1%	38	22,0%	2	13,3%

Anomie merena Sroleho škálou

Zde byl přidán řádek TOTAL (z nabídky STATISTICS)

Podíl osob s ruznou mírou anomie podle subjektivní třídy

	A75 TRIDA - SEBEZARAZENI SE							
	1 NIZSI TRIDA		2 NIZSI STREDNI		3 VYSSI STREDNI		4 VYSSI TRIDA	
	T01V0009 REKODOVANY INDEX ANOMIE		T01V0009 REKODOVANY INDEX ANOMIE		T01V0009 REKODOVANY INDEX ANOMIE		T01V0009 REKODOVANY INDEX ANOMIE	
	Count	%	Count	%	Count	%	Count	%
1 NIZKA	10	5,0%	107	18,0%	50	28,9%	7	46,7%
2 STREDNI	87	43,5%	260	43,8%	85	49,1%	6	40,0%
3 VYSOKA	103	51,5%	226	38,1%	38	22,0%	2	13,3%
$T Total	200	100,0%	593	100,0%	173	100,0%	15	100,0%

Anomie merena Sroleho škálou

ANALYZE

TABLES

TABLES OF FREQUENCIES

Frequencies for: ANOMREC
In each tables: A75 (subjektivní třída)
Separate tables: A98 (pohlaví)

Podíl osob s ruznou mírou anomie podle subjektivní třídy

A88 POHLAVI 1 MUZ

	A75 TRIDA - SEBEZARAZENI SE							
	1 NIZSI TRIDA		2 NIZSI STREDNI		3 VYSSI STREDNI		4 VYSSI TRIDA	
	T01V0009 REKODOVANY INDEX ANOMIE		T01V0009 REKODOVANY INDEX ANOMIE		T01V0009 REKODOVANY INDEX ANOMIE		T01V0009 REKODOVANY INDEX ANOMIE	
	Count	%	Count	%	Count	%	Count	%
1 NIZKA	5	2,5%	60	10,1%	32	18,5%	3	20,0%
2 STREDNI	36	18,0%	115	19,4%	48	27,7%	4	26,7%
3 VYSOKA	36	18,0%	112	18,9%	20	11,6%	2	13,3%
$T Total	77	38,5%	287	48,4%	100	57,8%	9	60,0%

Anomie merena Sroleho škálou

Objeví se jen tabulka pro 1. variantu znaku zadaného jako "separate tables" (zde pro muži). Klikneme-li 2x na tabulku, lze ji formátovat. Kliknutím na název proměnné vlevo nad tabulkou (A88 POHLAVI 1 MUZ) se objeví roletka s dalšími variantami (zde A88 POHLAVI 2 ZENA). Klikneme-li opět na ni, objeví se tabulka pro podsoubor ženy.

Podíl osob s ruznou mírou anomie podle subjektivní třídy

A88 POHLAVI 2 ZENA

	A75 TRIDA - SEBEZARAZENI SE							
	1 NIZSI TRIDA		2 NIZSI STREDNI		3 VYSSI STREDNI		4 VYSSI TRIDA	
	T01V0009 REKODOVANY INDEX ANOMIE		T01V0009 REKODOVANY INDEX ANOMIE		T01V0009 REKODOVANY INDEX ANOMIE		T01V0009 REKODOVANY INDEX ANOMIE	
	Count	%	Count	%	Count	%	Count	%
1 NIZKA	5	2,5%	47	7,9%	18	10,4%	4	26,7%
2 STREDNI	51	25,5%	145	24,5%	37	21,4%	2	13,3%
3 VYSOKA	67	33,5%	114	19,2%	18	10,4%	0	,0%
$T Total	123	61,5%	306	51,6%	73	42,2%	6	40,0%

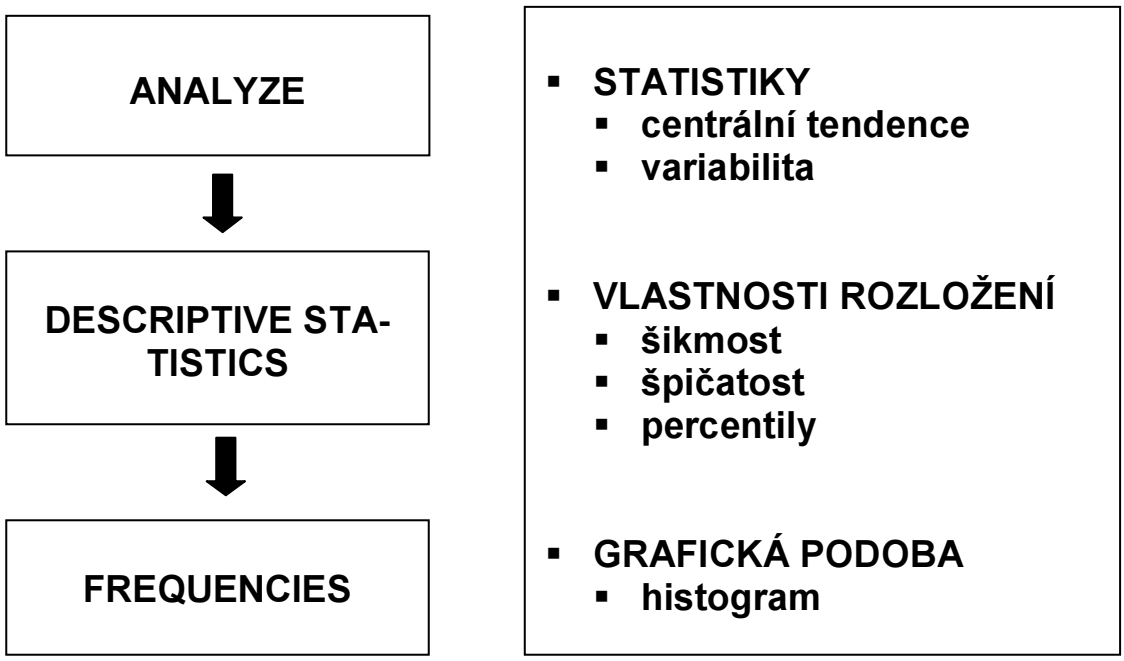
Anomie merena Sroleho škálou

LEKCE 02b
UNIVARIAČNÍ ANALÝZA SPOJITÝCH PROMĚNNÝCH

FREQUENCIES (KARDINÁLNÍ ZNAKY)

SPOJITÝ STATISTICKÝ ZNAK (kardinální):
- Nabývá všech možných hodnot z daného intervalu.
- V tomto případě se příliš nehovoří o četnosti určité hodnoty (je malá pravděpodobnost, že se stejná hodnota v souboru opakuje).
- I spojitý znak lze zobrazit a to stanovením intervalů, v nichž jsou určité hodnoty znaku (příjmové, věkové skupiny, ...).

Zobrazením není sloupcový diagram, ale HISTOGRAM. Jeho sloupce představují četnosti případů v intervalech.



MOŽNOSTI:
- Zadat i frekvenční tabulku (mají pouze omezený smysl) nebo jen statistiky
- Zadat statistiky (smysl mají všechny nabídnuté možnosti)
- Zadat diagram (smysl má histogram)
- Uspořádat data

FREQUENCIES - STATISTIKY

STŘEDNÍ HODNOTY (CENTRAL TENDENCY)

Sumární (typické) charakteristiky distribuce.

MODUS - (MODE)

Střední hodnota pro nominální znaky, ordinální a kardinální znaky: jde o kategorii s nejpočetnějším výskytem (obsahující nejvíce případů).

MEDIÁN – (MEDIAN)

Střední hodnota pro ordinální a kardinální znaky. Je to hodnota, dělící rozložení na dvě poloviny (50. percentil nebo též 2. kvartil). Někdy výhodnější než aritmetický průměr, neboť je rezistentní vůči extrémním hodnotám. U souborů, které mají lichý počet prvků je hodnota mediánu rovna hodnotě středního prvku. Při sudém počtu prvků se medián počítá jako aritmetický průměr hodnot dvou středních prvků. U ordinálních proměnných hovoříme o mediánové kategorii (hodnotě proměnné, v níž leží medián).

ARITMETICKÝ PRŮMĚR - (MEANS)

Střední hodnota pro kardinální znaky. Není vždy nejvhodnější - může se například značně změnit změnou i jen jednoho pozorování (citlivý na extrémní hodnoty).

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n) \qquad \text{neboli} \qquad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Řada statistických testů slouží k porovnávání průměrů, které získáme v různých podsouborech (sociálních kategoriích).

Příklad:
Porovnání průměrných platů u osob s různým dosaženým vzděláním.

Pro určité typy proměnných lze použít vždy jen určité střední charakteristiky. Pro:
- NOMINÁLNÍ proměnné jen modus.
- ORDINÁLNÍ proměnné modus a medián.
- KARDINÁLNÍ proměnné modus, medián, průměr).

PODOBA DISTRIBUCE DISTRIBUTION)

VARIABILITA (DISPERSION)

- MINIMUM - Je minimální hodnota rozdělení.
- MAXIMUM - Je maximální hodnota rozdělení.
- ROZPĚTÍ (RANGE) - Je rozdílem mezi nejvyšší (maximum) a nejnižší (minimum) hodnotou. Nejjednodušší míra variability, která nás upozorňuje na vzdálenost extrémních hodnot, ale nevyjadřuje vůbec koncentraci hodnot proměnné kolem středu rozložení.
- MEZIKVARTILOVÉ ROZPĚTÍ (IQR) - rozdíl mezi horním (75) a dolním (25) kvartilem. Lze ho použít (v kombinaci s ostatními charakteristikami) pro rozlišení toho jaká je variabilita (či koncentrace) hodnot proměnné kolem středu a na okrajích (v extrémních hodnotách) rozložení.
- ROZPTYL (VARIANCE) – Vypovídá o tom, jak jsou v rozložení hodnoty rozptýleny kolem aritmetického průměru. Je to průměrná čtvercová chyba (ve čtvercích jednotek původní proměnné) – součet druhých mocnin odchylek všech jednotlivých hodnot od průměru dělený rozsahem souboru.

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

- SMĚRODATNÁ/STANDARDNÍ ODCHYLKA (STDEV) Je druhou odmocninou rozptylu. Poskytuje míru hodnoty jakou má aritmetický průměr pro charakterizaci rozložení (čím je menší, tím lépe aritmetický průměr). Říká také, uvnitř jakého intervalu kolem průměru leží zvolené procento případů.

Rozptyl a směrodatná odchylka:
- Stejně jako průměr mají stejný rozměr jako měřená proměnná (například příjem: střední hodnota, rozptyl i směrodatní odchylka se vyjadřují v peněžní jednotce - v ČR v Kč, v UE v euro, v USA v dolarech ap.).
- Používají se jako kritéria toho jak moc se dá věřit či nevěřit průměru. Malé hodnoty rozptylu zvyšují význam průměru, velké znamenají, že hodnoty proměnné mají vysokou variabilitu a proto při používání průměru musíme být opatrní.
- Lze je použít jen pro porovnávání variability proměnných měřených ve stejných měrných jednotkách.

- KOEFICIENT VARIACE = $\frac{\text{standardní odchylka}}{\text{aritmetický průměr}} * 100$

Lze ho použít, na rozdíl od rozptylu a směrodatné odchylky, i pro porovnávání variability proměnných měřených v odlišných měrných jednotkách.

PERCENTILY

PERCENTIL (KVANTIL x_p)
Hodnota znaku, pro kterou platí, že nejméně p - procent případů má hodnotu menší nebo rovnu x_p a (100-p) případů je větších nebo rovno x_p.
Nejčastěji se používají:

- MEDIÁN neboli x_{50}
  50% případů má hodnotu menší než x_{50} a 50% větší než x_{50}.
- KVARTILY neboli x_{25}, x_{50}, x_{75} (nejčastěji dolní a horní kvartil).
  např. x_{25} = 25% případů má hodnotu menší než x_{25} a 75% větší než x_{25}.
- DECILY neboli x10, x20, x30, x40, x50, x60, x70, x80, x90.
  např. x_{20} = 20% případů má hodnotu menší než x_{20} a 80% větší než x_{20}.

Příklad použití:
- Jednou z kritérií pro určení chudoby je porovnání konkrétního příjmu s příjmovým rozložením. Například v EU je hranicí chudoby 60% mediánu příjmového rozložení (kdo má nižší příjem než je tato hranice, je považován za chudého).
- Může nás zajímat jak početný je spodní decil (nejchudší) a horní decil (nejbohatší) příjmového rozložení, ale i typické sociální charakteristiky osob ocitajících se ve spodním či horním decilu.

V obou případech, pokud přiřadíme každému jedinci nový znak (jak, to se dozvíme v bloku věnovaném transformaci proměnných), identifikující jeho polohu v takovémto rozložení - do kterého kvantilu svým příjmem patří, lze zkoumat strukturu tohoto kvantilu. Je například mezi osobami s příjmem pod hranicí chudoby (nebo ve spodním decilu) vyšší podíl osob s nějakou sociální charakteristikou (stupeň vzdělání, pohlaví, věk apod. než mezi ostatními osobami?

ŠIKMOST (SKEWNESS)

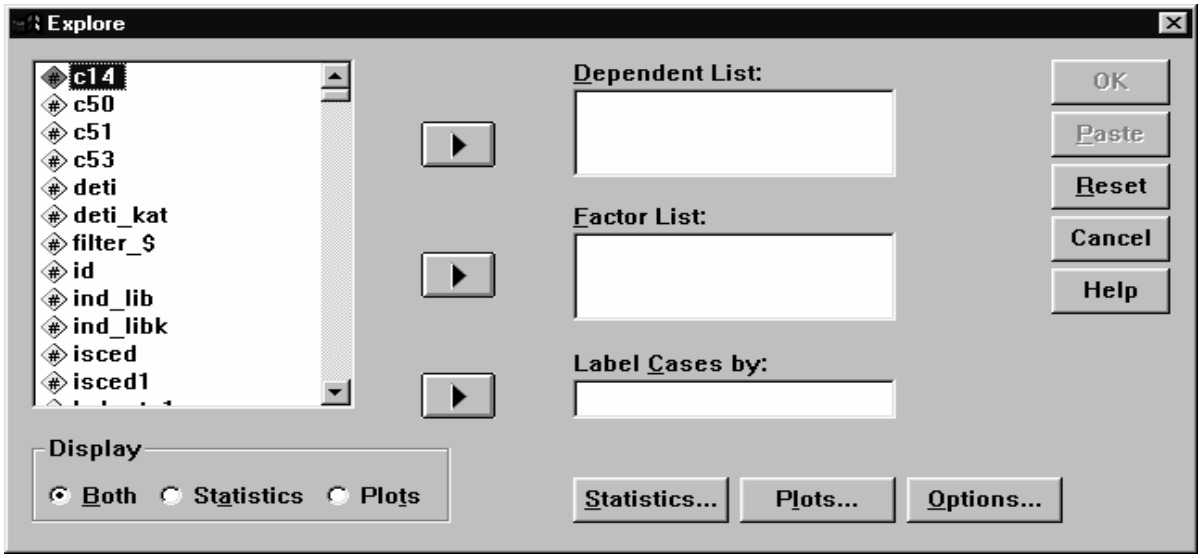
Charakteristiky šikmosti udávají, zda jsou hodnoty kolem zvoleného středu rozloženy souměrně, nebo je rozdělení hodnot zešikmeno na jednu stranu. Měří tedy asymetrii v distribuci hodnot:

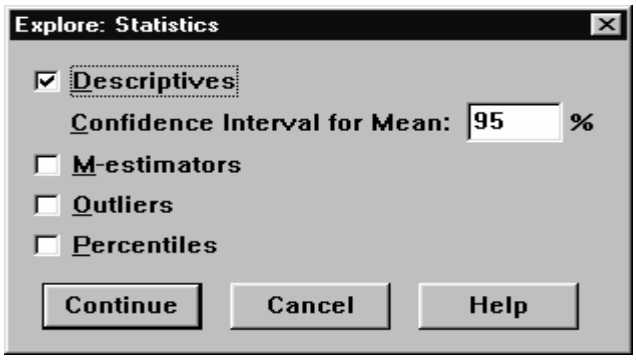
- 0 = symetrické rozložení (modus, medián, aritmetický průměr mají shodné či velmi blízké hodnoty).
- Kladná hodnota = šikmé doprava. Aritmetický průměr je větší než medián a ten je větší jako modus (více je případů menších než průměr).
- Záporná hodnota = šikmé doleva. Aritmetický průměr je menší než medián a ten je menší jako modus (více je případů větších než průměr)

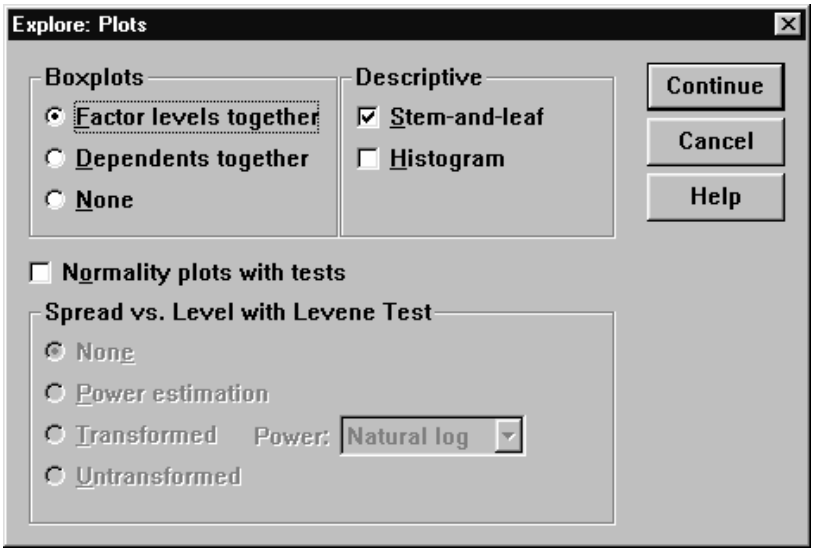
ŠPIČATOST (KURTOSIS)

Dána porovnáním s normálním rozložením. Čím je rozdělení špičatější, tím více jsou hodnoty soustředěny kolem jeho středu, čím je méně špičaté, tím častěji obsahuje hodnoty vzdálené od tohoto středu.
Kladná hodnota = více případů je mimo normální rozložení (plochá křivka).

PROCEDURA EXPLORE





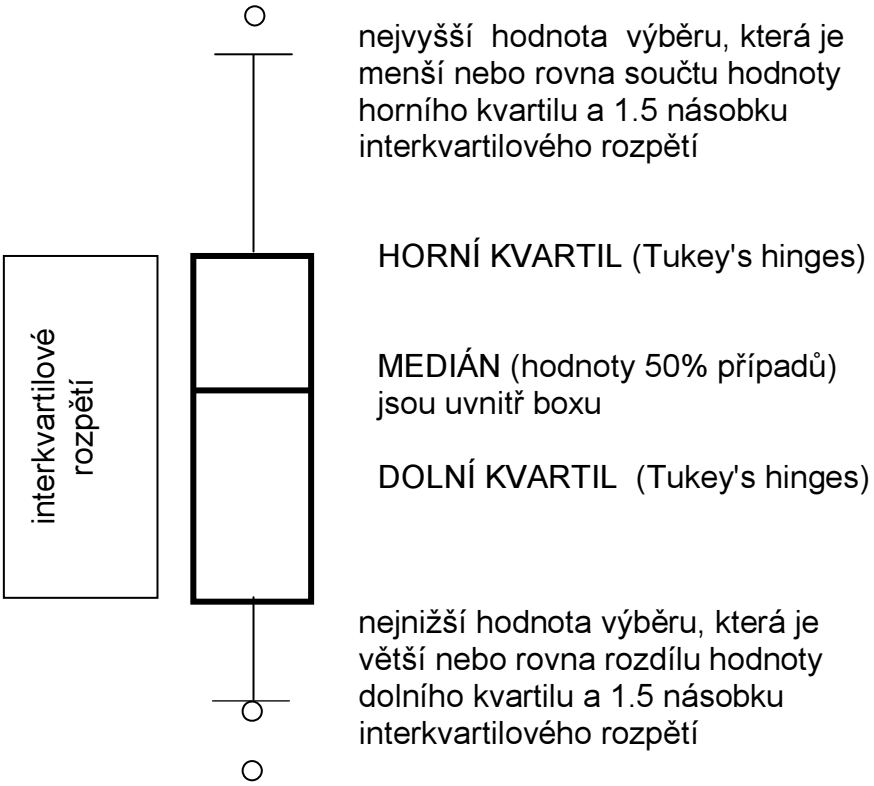


Co můžeme říci o datech podíváme-li se na BOXPLOT?

- Podle délky boxu můžeme určit šířku nebo variabilitu dat.
- Z mediánu můžeme určit centrální tendenci nebo polohu.
- Jestliže medián není uprostřed boxu můžeme usuzovat na sešikmení (skew).
  - Je-li medián blíže hornímu kvartilu jedná se o kladné sešikmení.
  - Je-li medián blíže dolnímu kvartilu jedná se o záporné sešikmení.

E - EXTREMES: hodnota vzdálená více než tři interkvartilová rozpětí od horního kvartilu

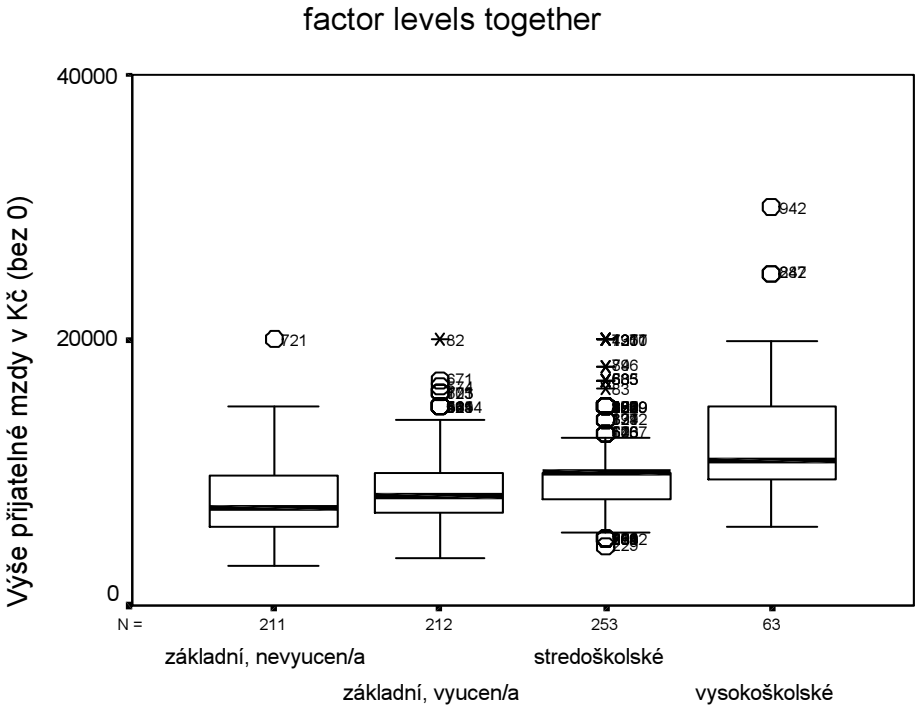
O - OUTLIERS: hodnota vzdálená více než 1.5 interkvartilového rozpětí od horního kvartilu



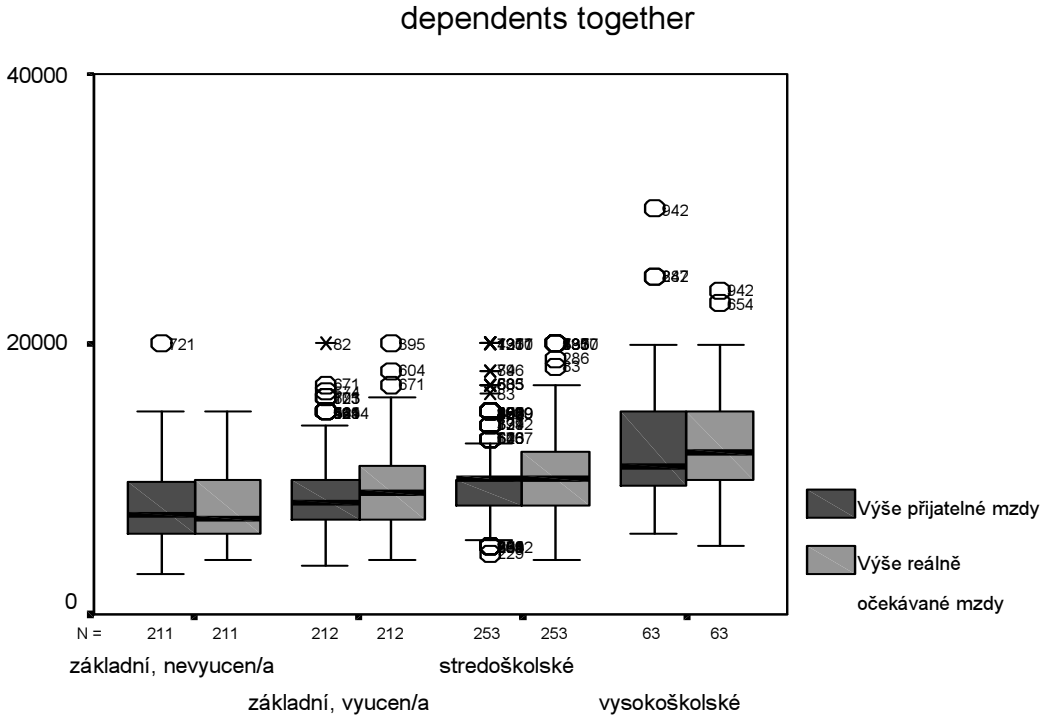
O - OUTLIERS: hodnota vzdálená více než 1.5 interkvartilového rozpětí od dolního kvartilu

E - EXTREMES: hodnota vzdálená více než tři interkvartilová rozpětí od dolního kvartilu

BOXPLOT je zvláště užitečný pro porovnávání hodnot v několika skupinách.

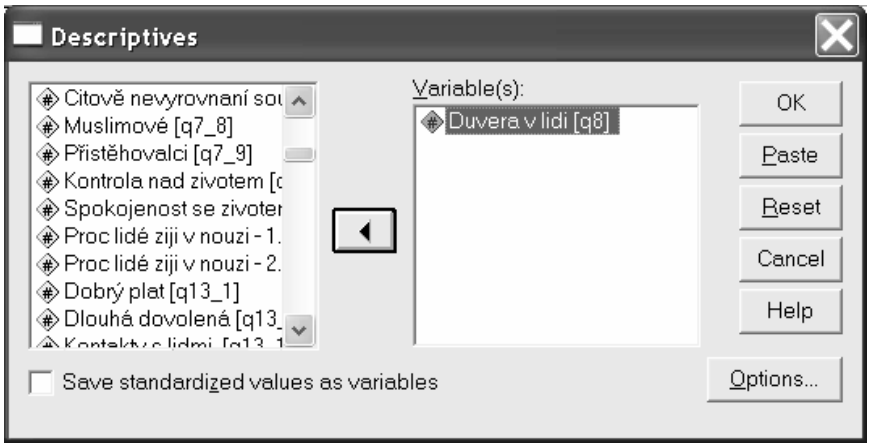


Nejvyšší dosažené vzdělání

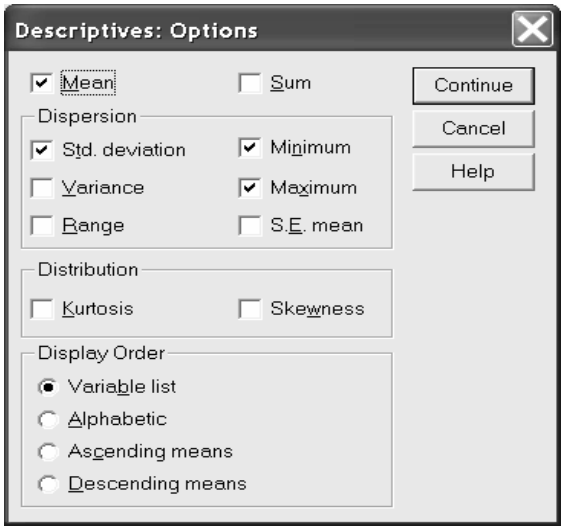


Nejvyšší dosažené vzdělání

DESCRIPTIVES



Tato procedura dává podobné výsledky jako FREQUENCIES či EXPLORE (spíše chudší). Například při volbě v OPTIONS:



To, co jsme zadali, dostaneme v následující tabulce:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Duvera v lidi	1869	1	2	1,76	,426
Valid N (listwise)	1869				

To ovšem není výsledek, který by nás velmi zajímal.

Zajímají nás však nově vytvořené z-skóre, respektive hodnoty této proměnné u jednotlivých případů. Příkazem save standardized values as variables vytvoříme totiž novou proměnnou nazvanou standardně jménem původní proměnné s předponou z (například vek → zvek). V matici je přidán sloupec s touto proměnnou (standardně je) a každému případu je přiřazena pro něj vypočítaná hodnota z-skóre. Tyto hodnoty nám říkají o kolik standardních odchylek a jakým směrem se každý z případů odchyluje od průměru rozložení dané proměnné (v tomto případě věk jednotky od věkového průměru souboru - blíže k tomu v příslušné lekci).