



Studijní materiál ke kurzu IBM IBM® SPSS® STATISTICS BASE: ZÁKLADY MODULU I.

Libor Šlik



2018

Strategický projekt UTB ve Zlíně, reg. č. CZ.02.2.69/0.0/0.0/16_015/0002204



EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MŠMT
MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



Informace o autorovi:

Libor Šlik, Ing.

ACREA CR, spol. s r.o.

lslik@acrea.cz



EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání





OBSAH

OBSAH.....	3
1 DESCRIPTIVES - POPISNÉ STATISTIKY	6
1.1 POPISNÉ STATISTIKY	6
1.2 VOLÁNÍ PROCEDURY V IBM SPSS STATISTICS	6
1.2.1 Nastavení dialogu Descriptives	6
1.2.2 Tlačítko Options	7
1.3 VÝSTUPY	8
1.3.1 Výstup procedury Descriptives.....	8
1.3.2 Výstup procedury Descriptives - seřazeno	9
2 FREQUENCIES - TABULKY ČETNOSTÍ	10
2.1 TABULKY ČETNOSTÍ	10
2.2 VOLÁNÍ PROCEDURY V IBM SPSS STATISTICS	10
2.2.1 Nastavení dialogu	10
2.2.2 Tlačítko Statistics.....	11
2.2.3 Tlačítko Charts.....	12
2.2.4 Tlačítko Format	13
2.3 VÝSTUPY	14
2.3.1 Statistika	14
2.3.2 Tabulka četností	14
2.3.3 Graf	15
3 EXPLORE - PODROBNÁ ANALÝZA DAT.....	16
3.1 VYUŽITÍ PROCEDURY EXPLORE.....	16
3.2 VOLÁNÍ PROCEDURY V IBM SPSS STATISTICS	16
3.2.1 Nastavení dialogu	16
3.2.2 Tlačítko Statistics.....	17
3.2.3 Tlačítko Plots.....	18
3.2.4 Tlačítko Options	19
3.3 VÝSTUPY	20
3.3.1 Přehled o počtu platných a chybějících případů.....	20
3.3.2 Popisné statistiky	20
3.3.3 Odhady střední hodnoty	21
3.3.4 Percentily	22
3.3.5 Výpis pěti nejnižších a nejvyšších hodnot.....	22
3.3.6 Boxplot	23
3.3.7 Histogram	24
3.3.8 Listový (cifrový) diagram	25
3.3.9 Graf pro ověřování normality dat (Q-Q plot)	26
3.3.10 Testy normality	27
3.3.11 Grafy závislosti variability na poloze	28
3.3.12 Test shody rozptylů.....	30





SEZNAM POUŽITÉ LITERATURY	31
SEZNAM OBRÁZKŮ	32
SEZNAM TABULEK	33





ÚVOD

Tento materiál slouží jako studijní materiál ke Strategickému projektu UTB ve Zlíně, reg. č. CZ.02.2.69/0.0/0.0/16_015/0002204.



EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání



1 DESCRIPTIVES - POPISNÉ STATISTIKY

1.1 Popisné statistiky

Popisné statistiky se užívají pro charakteristiku základních vlastností číselných proměnných (spojité znaky, počty výskytů nějaké události apod.).

Procedura Descriptives dále umožňuje uložení bezrozměrných standardizovaných proměnných (z-skórů), jejichž aritmetický průměr je nula a rozptyl má hodnotu jedna. Z-skóry se často používají například pro porovnání statistických vlastností proměnných, měřených v různých jednotkách. Výpočet z-skórů probíhá v každém řádku datové matice podle vzorce:

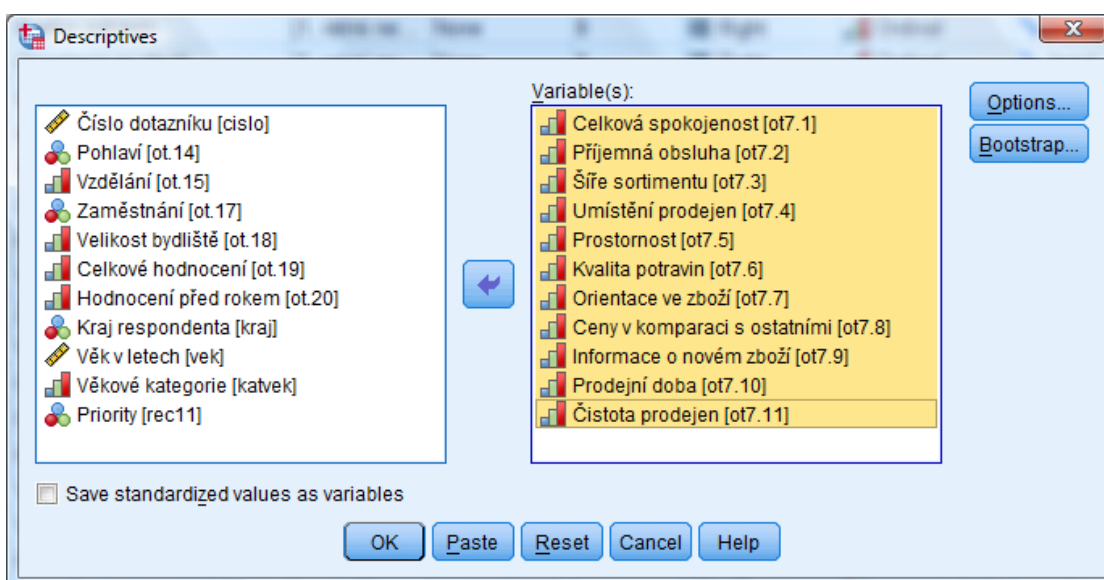
$$z_x = \frac{x - \bar{x}}{s_x},$$

kde $\bar{x}$ je aritmetický průměr proměnné X a s_x je její směrodatná odchylka.

1.2 Volání procedury v IBM SPSS Statistics

Analyze → Descriptive Statistics → Descriptives

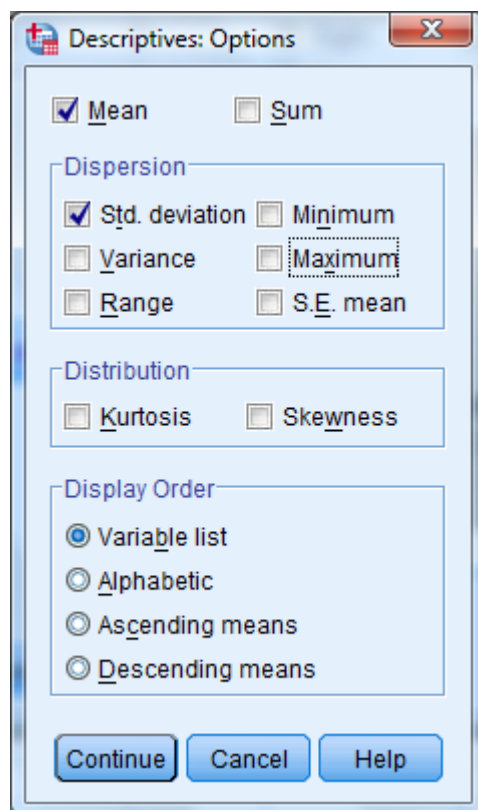
1.2.1 Nastavení dialogu Descriptives



Obrázek 1: Nastavení dialogu Descriptives

- Do pole Variable(s) přeneseme analyzované proměnné.
- Zaškrtneme-li Save standardized values as variables, vytvoří se na konci datové matice nové proměnné, obsahující z-skóry uvedených proměnných. Tyto proměnné můžeme využít k další analýze. Názvy nových proměnných se odvodí automaticky přidáním „Z“ před názvy původních proměnných.

1.2.2 Tlačítko Options



Obrázek 2: Nastavení Descriptives, tlačítko Options

Pomocí tlačítka *Options* volíme požadované statistiky a způsob uspořádání proměnných v tabulce. K dispozici jsou následující statistiky: *Mean* (aritmetický průměr), *Sum* (součet hodnot).

Dispersion (Míry variability)

Std. deviation (směrodatná odchylka), *Variance* (rozptyl), *Range* (rozpětí hodnot) - rozdíl maxima a minima, *Minimum* (minimum), *Maximum* (maximum), *S.E. mean* (standardní chyba aritmetického průměru).

Distribution (Rozdělení)

Kurtosis (špičatost), *Skewness* (šikmost).

Display order (Řazení proměnných)

Proměnné v tabulce mohou být uspořádány těmito způsoby:

Variable list (podle pořadí v souboru), *Alphabetic* (podle abecedy), *Ascending means* (podle rostoucích průměrů), *Descending means* (podle klesajících průměrů).

1.3 Výstupy

Výstupem procedury je tabulka zvolených statistik a případně také nové proměnné obsahující z-skóry.

1.3.1 Výstup procedury Descriptives

Descriptive Statistics			
	N	Mean	Std. Deviation
Celková spokojenost	505	3.17	1.558
Příjemná obsluha	505	2.91	1.589
Šíře sortimentu	504	3.36	1.375
Umístění prodejen	503	4.36	1.500
Prostornost	503	2.60	1.467
Kvalita potravin	503	2.86	1.384
Orientace ve zboží	502	3.26	1.553
Ceny v komparaci s ostatními	503	2.64	1.397
Informace o novém zboží	501	3.54	1.486
Prodejní doba	504	3.27	1.471
Čistota prodejen	502	3.99	1.556
Valid N (listwise)	497		

Tabulka 1 Výstup procedury Descriptives



Jestliže zpracováváme například sady proměnných hodnocených na stejné škále, tabulku velmi zpřehledníme seřazením proměnných podle průměrné hodnoty (Viz Obrázek 2: Nastavení Descriptives, tlačítko Options).

1.3.2 Výstup procedury Descriptives - seřazeno

Na následujícím obrázku jsou proměnné v tabulce seřazeny nejprve podle pořadí v souboru a poté podle průměrné spokojenosti (od nejvyšší po nejnižší).

Descriptive Statistics			
	N	Mean	Std. Deviation
Umístění prodejen	503	4.36	1.500
Čistota prodejen	502	3.99	1.556
Informace o novém zboží	501	3.54	1.486
Šíře sortimentu	504	3.36	1.375
Prodejní doba	504	3.27	1.471
Orientace ve zboží	502	3.26	1.553
Celková spokojenost	505	3.17	1.558
Příjemná obsluha	505	2.91	1.589
Kvalita potravin	503	2.86	1.384
Ceny v komparaci s ostatními	503	2.64	1.397
Prostornost	503	2.60	1.467
Valid N (listwise)	497		

Tabulka 2 Výstup procedury Descriptives - seřazeno



2 FREQUENCIES - TABULKY ČETNOSTÍ

2.1 Tabulky četností

Z tabulek četností získáme základní informaci o rozložení hodnot kategorizovaných proměnných. Tento typ tabulky naopak není vhodný pro spojité číselné proměnné vzhledem k tomu, že nabývají mnoha různých hodnot a každá z nich se dokonce může v datech vyskytovat pouze jednou. Výstupní tabulka obvykle obsahuje:

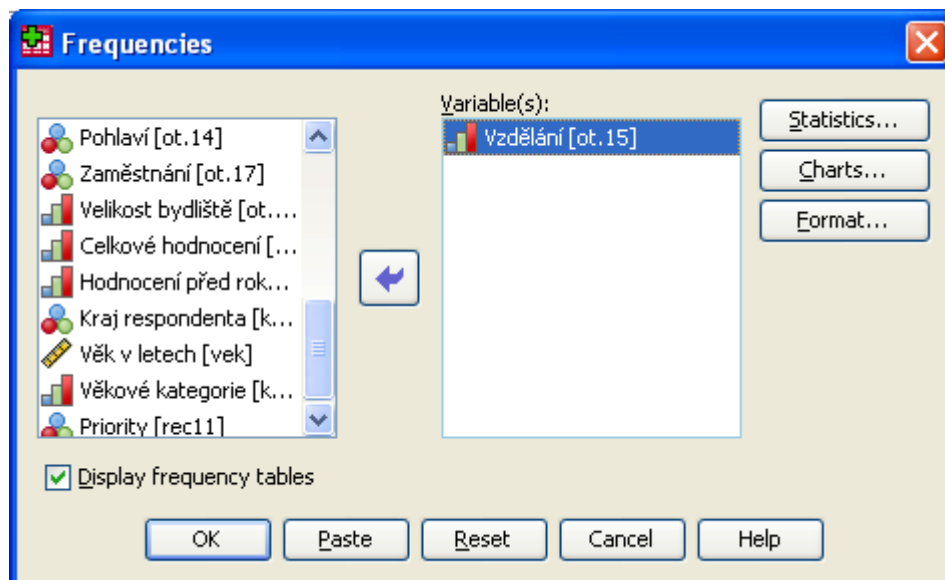
- Četnosti (rozdělené na platné a chybějící hodnoty)
- Procenta (vzhledem k celému souboru i pouze k platným hodnotám)
- Kumulativní procenta

Tabulku lze doplnit grafem, nejčastěji sloupcovým nebo koláčovým, nebo dalšími statistikami.

2.2 Volání procedury v IBM SPSS Statistics

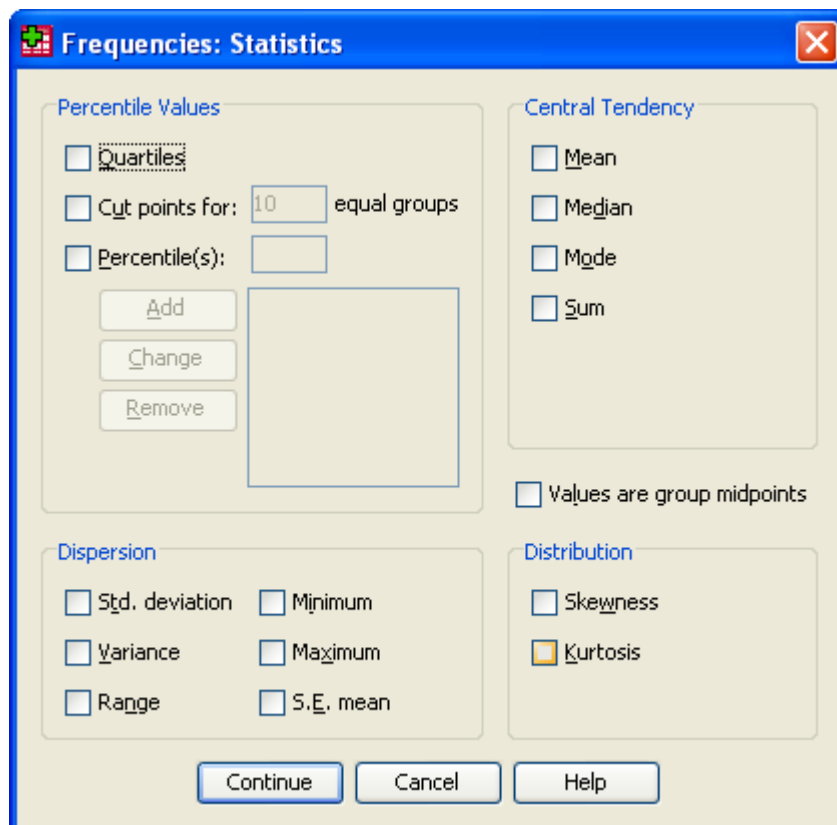
Analyze → Descriptive Statistics → Frequencies

2.2.1 Nastavení dialogu



Obrázek 3: Nastavení dialogu Frequencies

2.2.2 Tlačítko Statistics



Obrázek 4: Nastavení dialogu Frequencies, tlačítko Statistics

Percentile Values (kvantily a percentily)

Kvantily dělí množinu hodnot na stejně početné části. Percentily vyjadřují hodnotu, která odděluje dané procento nižších hodnot od vyšších.

- Quartiles (kvartily) - dělí množinu hodnot na čtyři stejně početné části.
- Cut points for ... equal groups (obecné kvantily) - do pole zadáme požadovaný počet částí, například 10 odpovídá decilům.
- Percentile(s) (percentily) - tlačítkem Add přidáváme požadované percentily do seznamu. Zadávané hodnoty musí být z intervalu $<0;100>$.

Central tendency (Centrální tendence)

Následující statistiky vyjadřují polohu rozdělení proměnné:

Mean (aritmetický průměr), Median (medián) - dělí množinu hodnot na dvě stejně početné části, Mode (modus) - nejčastější hodnota, Sum (součet).

Values are group midpoints (Hodnoty jsou středy intervalů)

Jestliže analyzovaná proměnná vznikla kategorizací spojité proměnné a hodnoty proměnné odpovídají středům intervalů, označením této volby získáte odhady mediánu a percentilů pro původní spojitou proměnnou.

Dispersion (Rozptýlení):

Statistiky vyjadřující variabilitu a rozptýlení proměnné:

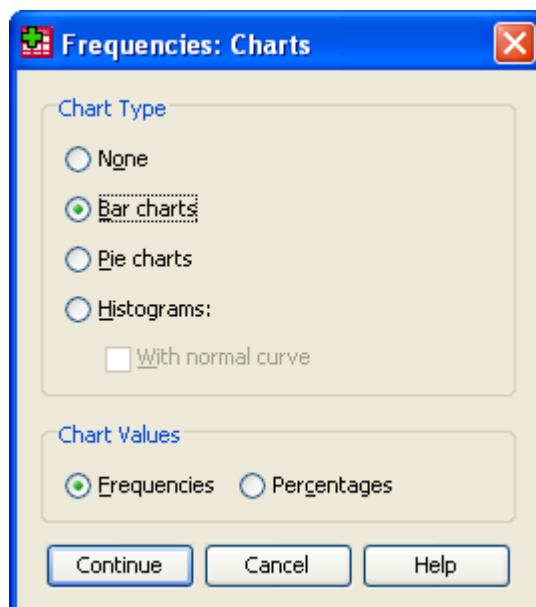
Std. deviation (směrodatná odchylka), Variance (rozptyl), Range (rozpětí hodnot), Minimum (minimum), Maximum (maximum), S. E. mean (standardní chyba průměru).

Distribution (Rozdělení):

Následující statistiky informují o symetrii a tvaru rozložení proměnné:

Skewness (šikmost), Kurtosis (špičatost).

2.2.3 Tlačítko Charts



Obrázek 5: Nastavení dialogu Frequencies, tlačítko Charts

Chart Type (Typ grafu)

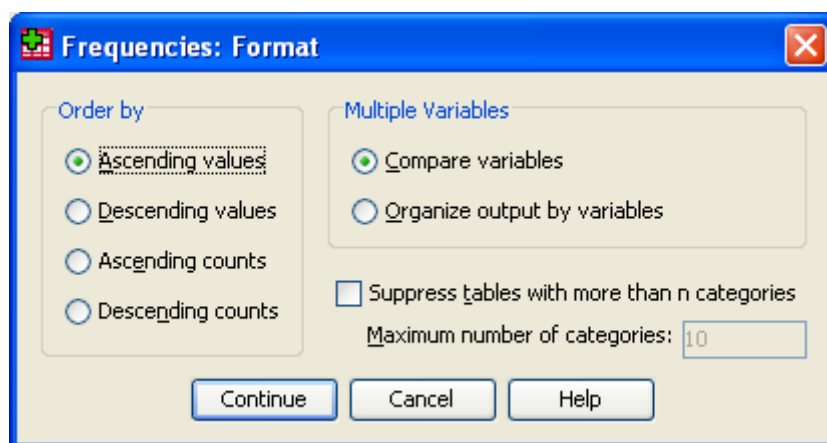
Tabulku četností lze doplnit následujícími typy grafů:

None (žádný graf), Bar charts (sloupcový graf), Pie charts (koláčový graf) a Histogram (histogram), případně s křivkou hustoty normálního rozdělení.

Chart Values (Hodnoty grafu)

Dále nastavíme, zda se v grafech mají zobrazovat četnosti (Frequencies) nebo procenta (Percentages).

2.2.4 Tlačítko Format



Obrázek 6: Nastavení dialogu Frequencies, tlačítko Format

Order by (Setřídění výstupní tabulky)

Výstupní tabulka četností může být setříděna podle:

rostoucích hodnot (Ascending values), klesajících hodnot (Descending values), rostoucích četností (Ascending counts), klesajících četností (Descending counts).

Multiple Variables (Více proměnných)

Jestliže analyzujeme současně více proměnných, rozhodujeme v této části dialogového okna, jakým způsobem organizovat výstup:

- Compare variables (srovnání proměnných) - ve výstupovém okně se zobrazí nejprve jeden typ výstupu (například tabulka četností) pro všechny analyzované proměnné a dále následuje další typ výstupu (například graf) opět pro všechny proměnné.
- Organize output by variables (organizace výstupu po proměnných) - nejprve všechny výstupy vztahující se k první proměnné a dále následují analýzy dalších proměnných.

Suppress tables with more than n categories (Potlačení velkých tabulek)

V této části dialogového okna můžeme nastavit potlačení tabulek s příliš velkým počtem kategorií. Tabulky četností pro proměnné, které mají větší, než uvedený počet kategorií, se potom nebudou ve výstupním okně zobrazovat.

2.3 Výstupy

2.3.1 Statistiky

Informace o platných a chybějících hodnotách. Tabulka může být doplněna také dalšími statistikami, pokud byly požadovány.

Statistics		
Vzdělání		
N	Valid	503
	Missing	2

Tabulka 3 Výstup procedury Frequencies, Statistics

2.3.2 Tabulka četností

Kategorie tabulky četností jsou rozděleny na platné a chybějící. Ve sloupcích jsou postupně: četnosti (Frequency), procenta (Percent), procenta z platných případů (Valid Percent), kumulativní procenta (Cumulative Percent). Poslední sloupec má smysl pouze u ordinálních proměnných nebo v případě, že jsou kategorie smysluplným způsobem seřazeny.

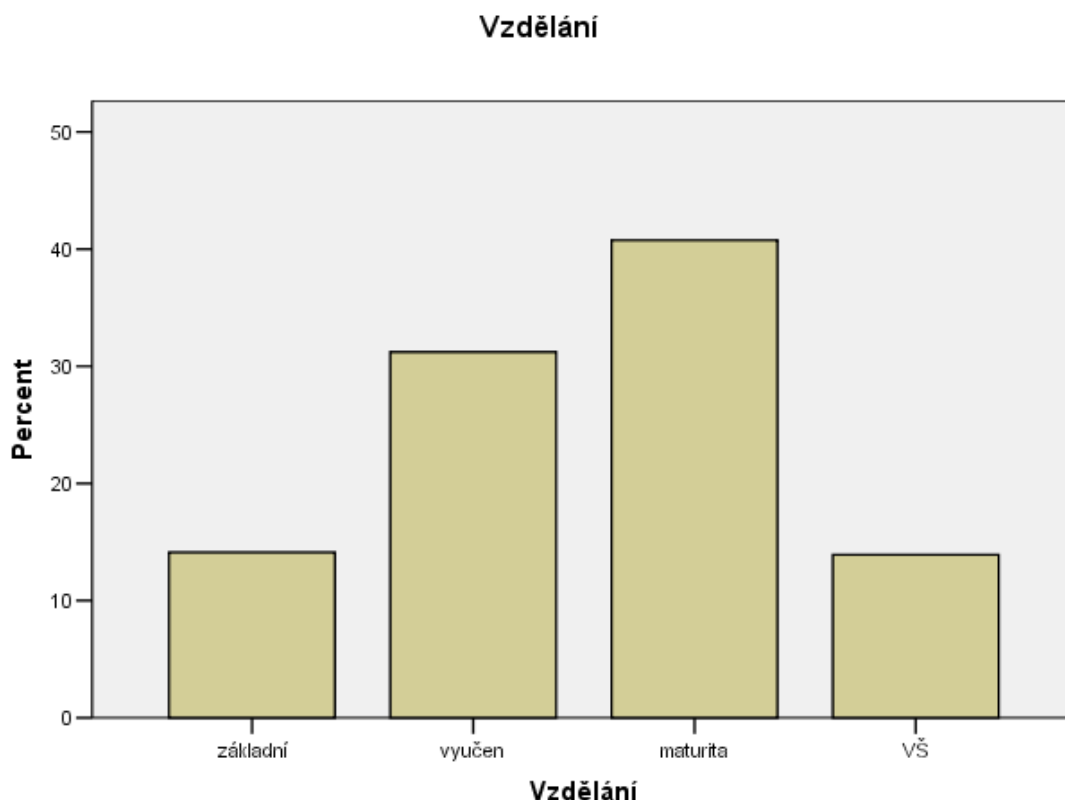


Vzdělání					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	základní	71	14.1	14.1	14.1
	vyučen	157	31.1	31.2	45.3
	maturita	205	40.6	40.8	86.1
	VŠ	70	13.9	13.9	100.0
	Total	503	99.6	100.0	
Missing	System	2	.4		
Total		505	100.0		

Tabulka 4 Výstup procedury Frequencies, tabulka četností

2.3.3 Graf

Zvolený typ grafu. Pro ordinální proměnné obvykle užíváme sloupcový graf, který graficky zdůrazňuje uspořádání kategorií. Pro nominální proměnné se velmi dobře hodí koláčový graf.



Obrázek 7: Výstup procedury Frequencies, graf

3 EXPLORE - PODROBNÁ ANALÝZA DAT

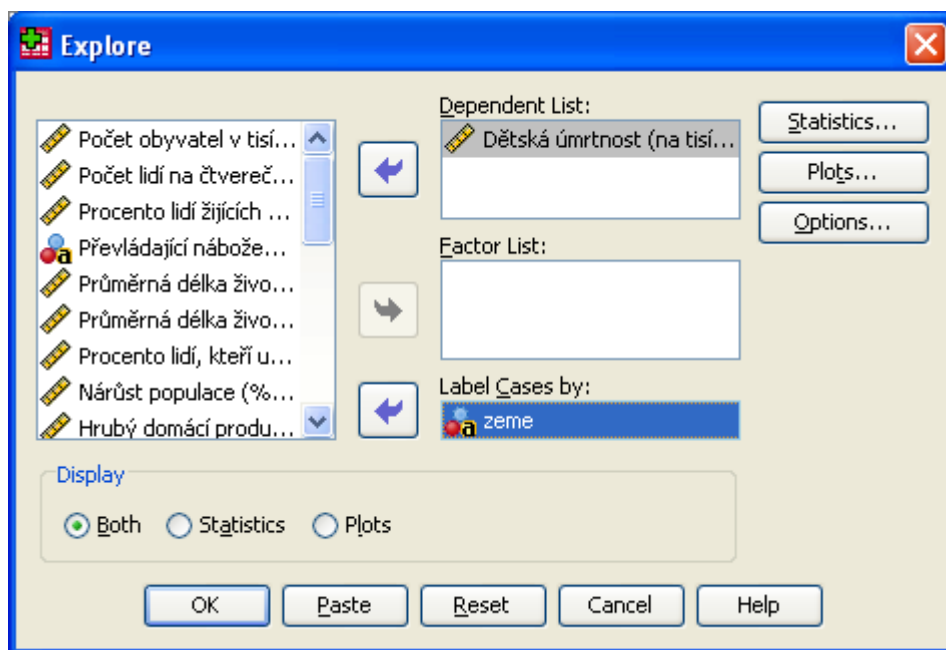
3.1 Využití procedury Explore

Procedura Explore je určena k podrobné analýze číselných proměnných. Je zde k dispozici celá řada popisných statistik, popisné statistické grafy a také testování normality a homogenity variancí. Datový soubor analyzujeme vcelku nebo v určených skupinách (definovaných na základě kategorizovaných proměnných - faktorů), které charakterizujeme zvlášť.

3.2 Volání procedury v IBM SPSS Statistics

Analyze → Descriptive Statistics → Explore

3.2.1 Nastavení dialogu



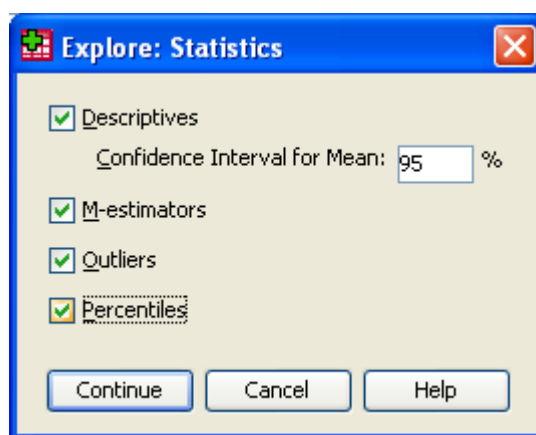
Obrázek 8: Nastavení dialogu Explore

- Do okna Dependent List přeneseme analyzované číselné proměnné.
- V poli Factor List zadáváme jednu nebo více kategorizovaných proměnných, které rozdělí datový soubor na skupiny. Tyto skupiny soubor v analýze člení

a jsou buď samostatně a/nebo komparačně (graficky) popisovány. V případě, že zadáme více faktorů, provedou se analýzy pro skupiny určené každým z nich zvlášť.

- Do okna Label Cases by umísťujeme proměnnou, která určuje názvy jednotlivých případů.
- V části okna označené Display rozhodujeme, zda nás zajímají pouze statistiky, pouze grafy nebo obojí.

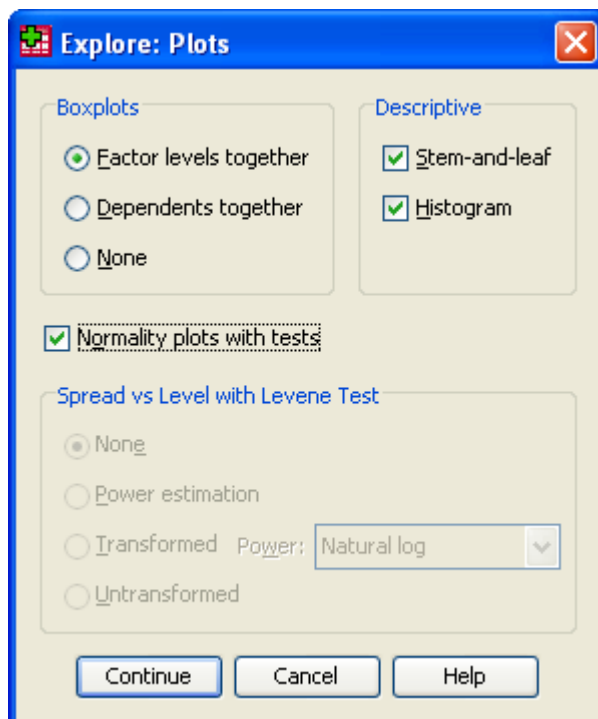
3.2.2 Tlačítko Statistics



Obrázek 9: Nastavení dialogu Explore, tlačítko Statistics

- Při zaškrtnutí políčka *Descriptives* se vytvoří tabulka popisných statistik. Na rozdíl od většiny procedur IBM SPSS Statistics zde není třeba podrobněji zadávat požadované statistiky, automaticky se vytvoří tabulka se standardním přehledem. Pro výpočet intervalu spolehlivosti pro průměr lze určit požadovanou hladinu spolehlivosti (*Confidence Interval for Mean*).
- Políčko *M-estimators* ovlivňuje výpočet dalších statistik pro odhad střední hodnoty.
- Pro výpis pěti nejvyšších a pěti nejnižších hodnot je určeno políčko *Outliers*.
- Označíme-li *Percentiles*, zobrazí se tabulky s percentily zkoumaných proměnných.

3.2.3 Tlačítko Plots



Obrázek 10: Nastavení dialogu Explore, tlačítko Plots

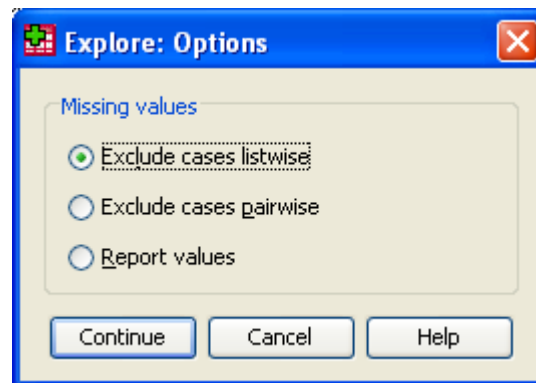
- Přepínač *Boxplots* určuje zda, a případně jakým způsobem, se vytvoří boxplot. Pro každou zkoumanou proměnnou lze nakreslit samostatný diagram (*Factor levels together*), ve kterém budou grafy pro jednotlivé skupiny, nebo můžeme vytvořit jediný graf pro porovnávání vlastností všech analyzovaných proměnných (*Dependents together*). Při volbě *None* se žádný boxplot nevytvoří.
- Listový diagram vytvoříme označením políčka *Stem-and-leaf*.
- Zaškrťovací políčko *Histogram* řídí zobrazení histogramu.
- Grafy a výstupní tabulky testů normality získáme zaškrtnutím *Normality plots with tests*.

V části *Spread vs. Level with Levene Test* zadáváme grafy pro závislost heterogeneity na poloze. Tyto grafy jsou k dispozici pouze tehdy, je-li zadán nějaký faktor.

- Pro vykreslení netransformovaného grafu zvolíme *Untransformed* - graf potom zobrazuje vztah mediánu a kvartilového rozpětí ve skupinách.

- Označíme-li *Power estimation*, zobrazí se graf s logaritmickými škálami na osách. Dále je doporučena transformace typu obecná mocnina ($Y = a \cdot X^p$), která co nejlépe odstraní korelaci mezi variabilitou a polohou skupin.
- Při volbě *Transformed* dále vybereme z rozbalovacího seznamu *Power* některou z možných transformací před výpočtem mediánu a kvartilového rozpětí. K dispozici jsou tyto transformace: přirozený logaritmus (Natural log), převrácená hodnota druhé odmocniny (1/square root), převrácená hodnota (Reciprocal), druhá odmocnina (Square root), druhá mocnina (Square), třetí mocnina (Cube).
- Pokud nechceme zobrazit žádný z těchto grafů, označíme *None*.

3.2.4 Tlačítko Options



Obrázek 11: Nastavení dialogu Explore, tlačítko Options

Přepínačem Missing Values určíme, jak zacházet s chybějícími hodnotami.

- Při volbě *Exclude cases listwise* program vyloučí ze všech analýz případy, které mají u některé z analyzovaných proměnných vynechanou hodnotu. Všechny analýzy tedy vycházejí ze stejných případů, pokud je však chybějících hodnot větší počet, může se tímto způsobem počet analyzovaných případů výrazně snížit.
- Označíme-li *Exclude cases pairwise*, vyloučí se v každé analýze samostatně pouze nezbytně nutné případy. Výhodou tohoto přístupu je využití maximálního možného počtu případů. Nevýhodou mohou být naopak různé počty případů, ze kterých jednotlivé analýzy vycházejí.

- Jestliže také segmentační proměnná obsahuje vynechané hodnoty, lze na jejich základě definovat samostatnou skupinu (Report values).

3.3 Výstupy

3.3.1 Přehled o počtu platných a chybějících případů

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Dětská úmrtnost (na tisíc živých porodů)	109	100.0%	0	.0%	109	100.0%

Tabulka 5 Přehled o počtu platných a chybějících případů

3.3.2 Popisné statistiky

Descriptives					Statistic	Std. Error
Dětská úmrtnost (na tisíc živých porodů)	Mean				42.313	3.6473
	95% Confidence Interval for Mean	Lower Bound			35.083	
		Upper Bound			49.543	
	5% Trimmed Mean				39.341	
	Median				27.700	
	Variance				1450.027	
	Std. Deviation				38.0792	
	Minimum				4.0	
	Maximum				168.0	
	Range				164.0	
	Interquartile Range				55.3	
	Skewness				1.090	.231
	Kurtosis				.365	.459

Tabulka 6 Popisné statistiky z procedury Explore

Tabulka popisných statistik obsahuje přehled základních charakteristik dat: průměr (*Mean*), dolní a horní mez intervalu spolehlivosti pro průměr (*Confidence Interval for Mean*), 5% useknutý průměr, tj průměr po vyloučení 5 % nejvyšších a 5 %

nejnižších hodnot (5% *Trimmed Mean*), medián (*Median*), rozptyl (*Variance*), směrodatnou odchylku (*Std. Deviation*), minimum (*Minimum*), maximum (*Maximum*), rozpětí (*Range*), kvartilové rozpětí (*Interquartile Range*), šikmost (*Skewness*) a špičatost (*Kurtosis*). Dále jsou uvedeny také standardní chyby některých veličin.

3.3.3 Odhady střední hodnoty

M-Estimators				
	Huber's M-Estimator ^a	Tukey's Biweight ^b	Hampel's M-Estimator ^c	Andrews' Wave ^d
Dětská úmrtnost (na tisíc živých porodů)	32.694	28.735	34.296	28.473

a. The weighting constant is 1.339.

b. The weighting constant is 4.685.

c. The weighting constants are 1.700, 3.400, and 8.500

d. The weighting constant is $1.340 \cdot \pi$.

Tabulka 7 Odhady střední hodnoty

Odhady střední hodnoty zobrazené v tabulce *M-Estimators* dovolují uživateli posoudit, do jaké míry je průměrná hodnota v tomto případě ovlivněna extrémny, a nabízejí možnost zvolit vhodnou statistiku pro odhad střední hodnoty. Vzhledem k tomu, že aritmetický průměr je velmi citlivý na odlehlá pozorování, je často vhodné tyto extrémny eliminovat. Extrémny hodnoty se tedy mohou z výpočtu vyřadit úplně, nebo jim lze přiřadit menší váhu. Uvedené statistiky jsou modifikací běžného aritmetického průměru, vycházejí však z myšlenky, že čím dále se hodnota nachází od centra dat, tím menší váhu by ve výpočtu měla mít. Nabízené odhady (*Huberův*, *Tukeyův*, *Hampelův* a *Andrewsův*) kombinují různými způsoby tento přístup. Jedná se o iterační metody, které pracují se standardizovanými hodnotami (z-skóry) a přidělují datům váhy na základě porovnání s určitými konstantami. Konkrétní popis metod je k dispozici v nápovědě IBM SPSS Statistics.

3.3.4 Percentily

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	Dětská úmrtnost (na tisíc živých porodů)	5.700	6.500	9.250	27.700	64.500	110.000	117.500
Tukey's Hinges	Dětská úmrtnost (na tisíc živých porodů)			9.300	27.700	63.000		

Tabulka 8 Percentily

Percentily umožňují mapovat rozdělení proměnných. Kromě mediánu a kvartilů se zobrazí také percentily pro $p = 5\%$, 10% , 90% a 95% . Tabulka obsahuje nejen hodnoty standardních percentilů, ale i hlavní kvantily počítané Tukeyho metodou, které se nepatrně liší od standardního výpočtu (používá se v boxplotu).

3.3.5 Výpis pěti nejnižších a nejvyšších hodnot

Extreme Values					Case Number	zeme	Value
Dětská úmrtnost (na tisíc živých porodů) Highest					1	Afganistán	168.0
					22	JAR	137.0
					90	Somálsko	126.0
					40	Gambie	124.0
					17	Burkina Faso	118.0
Lowest					49	Island	4.0
					57	Japonsko	4.4
					96	Taiwan	5.1
					37	Finsko	5.3
					93	Švédsko	5.7 ^a

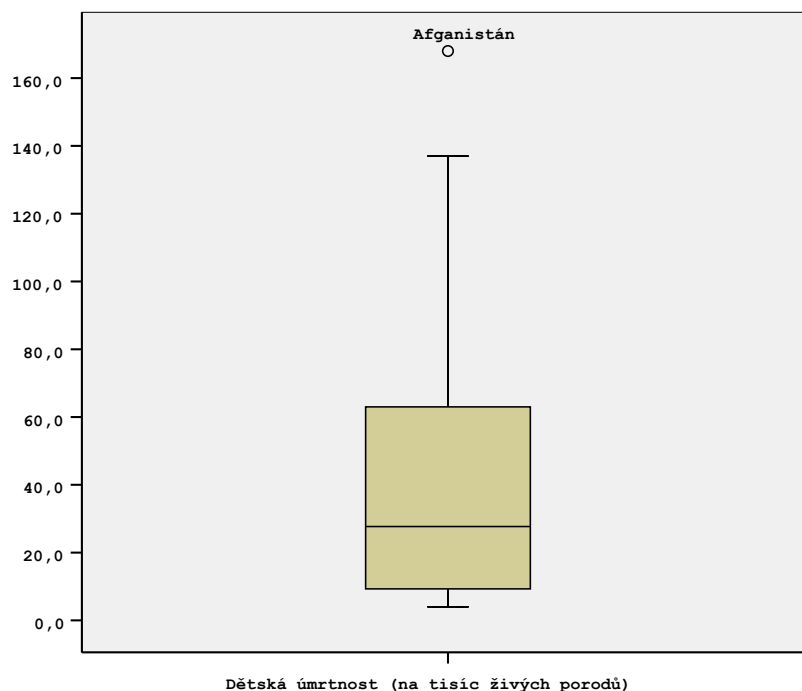
a. Only a partial list of cases with the value 5.7 are shown in the table of lower extremes.

Tabulka 9 Extreme Values

Výpis pěti nejnižších a nejvyšších hodnot umožňuje na základě znalosti extrémních hodnot odhalit chyby v měření či zadávání dat. IBM SPSS Statistics zobrazuje pět nejvyšších a pět nejnižších hodnot pro každý datový segment.

Poznámka pod tabulkou informuje, že hodnota 5.7 se v datech vyskytuje vícekrát.

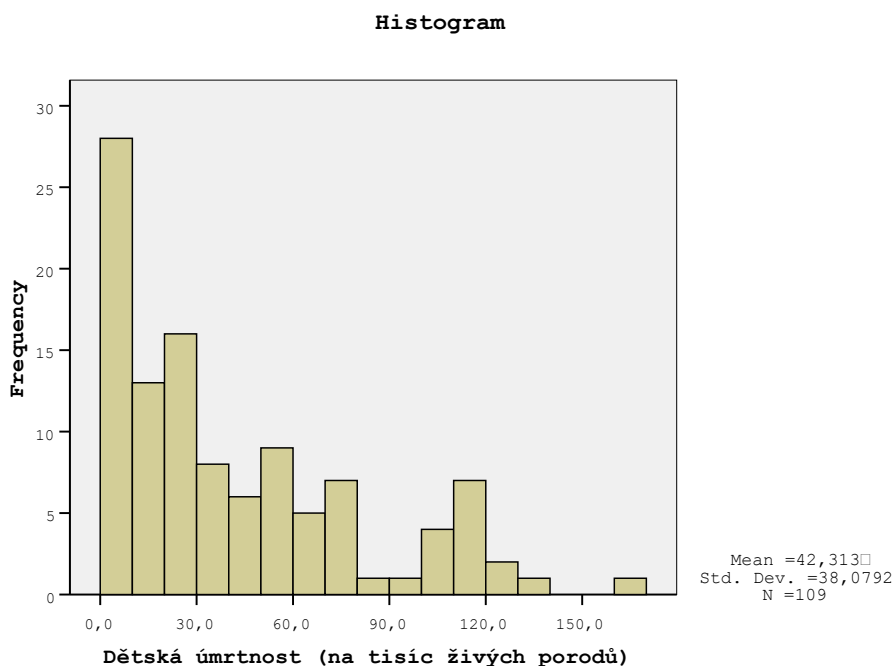
3.3.6 Boxplot



Obrázek 12 Boxplot

Boxplot (krabicový graf) poskytuje základní informaci o rozložení proměnné a názorně zachycuje medián, kvartily i extrémy v datech. Dolní a horní okraj krabičky odpovídají kvartilům, dělicí čára uvnitř krabičky vyjadřuje medián. Jestliže se všechny případy nacházejí maximálně ve vzdálenosti 1,5-násobku kvartilového rozpětí od nejbližšího okraje krabičky, naznačuje dolní výběžek minimální hodnotu a horní výběžek maximální hodnotu. Pokud je však hranice 1,5-násobku kvartilového rozpětí překročena, jsou tyto případy vyznačeny v grafu samostatně jako extrémy nebo odlehlá pozorování a výběžky v grafu jsou vyznačeny bez ohledu na tyto případy. Za odlehlá pozorování se přitom považují ty hodnoty, které leží ve vzdálenosti od 1,5-násobku do 3-násobku kvartilového rozpětí (v grafu se značí kolečkem), za extrémy hodnoty, které překročily hranici 3-násobku kvartilového rozpětí (v grafu se značí hvězdičkou).

3.3.7 Histogram



Obrázek 13 Histogram

Histogram je dalším typem grafu, který informuje o rozložení sledované proměnné. Hodnoty na ose X jsou rozděleny do stejně širokých intervalů a ve sloupcích jsou vyneseny četnosti těchto intervalů. Počet intervalů je určen optimálně pomocí algoritmu, při editaci však lze upravit. V pravé části grafu jsou dále zobrazeny hodnoty průměru, směrodatné odchylky a počet případů.

3.3.8 Listový (cifrový) diagram

Dětská úmrtnost (na tisíc živých porodů) Stem-and-Leaf Plot

Frequency	Stem &	Leaf
28.00	0 .	4455555666666666777778888899
13.00	1 .	0122223467799
16.00	2 .	0001123555577788
8.00	3 .	45567999
6.00	4 .	135679
9.00	5 .	011222347
5.00	6 .	03678
7.00	7 .	4556679
1.00	8 .	5
1.00	9 .	4
4.00	10 .	1569
7.00	11 .	0022378
2.00	12 .	46
1.00	13 .	7
1.00	Extremes	(>=168)

Stem width: 10.0
Each leaf: 1 case(s)

Obrázek 14 Stem&Leaf graf

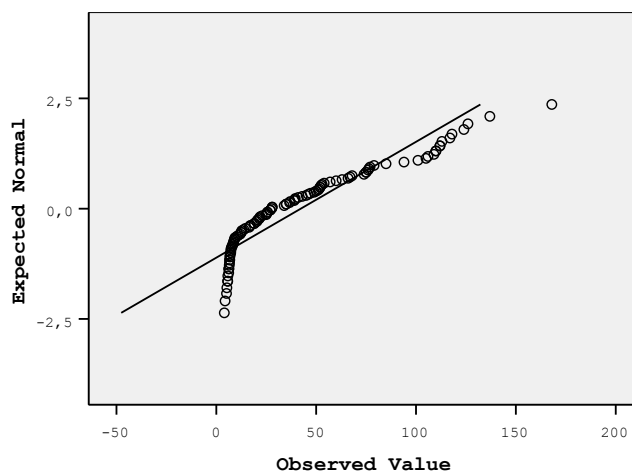
Listový (cifrový) diagram (Stem and leaf) je obdobou histogramu, proti histogramu se však kreslí otočený o 90°. Diagram současně charakterizuje četnosti jednotlivých intervalů i zobrazuje konkrétní data. Hodnoty jsou rozděleny na základ (stem), který vyjadřuje několik prvních cifer číselného zápisu, a zbytek (leaf), vyjádřený následující cifrou. Ve sloupci *Frequency* jsou zobrazeny četnosti kategorií základu. V části *Leaf* se vypíše, kolikrát se zde vyskytovaly jednotlivé následující cifry. Přitom jeden zápis čísla může zastupovat více případů, což je případně poznamenáno pod grafem. V tomto případě se tedy například základ 12 vyskytoval celkem dvakrát, konkrétně se přitom jednalo o hodnoty 124 a 126.

Výhodou cifrového diagramu proti histogramu je větší množství obsažené informace - data jsou seříděna a seskupena a známe nejen četnosti intervalů, ale také všechny originální hodnoty. Graf lze rovněž velmi snadno vytvářet i ručně a průběžně doplňovat další hodnoty tak, jak jsou sbírány, což bylo dříve velkou výhodou. Tento typ grafu je vhodný především pro menší datové soubory.



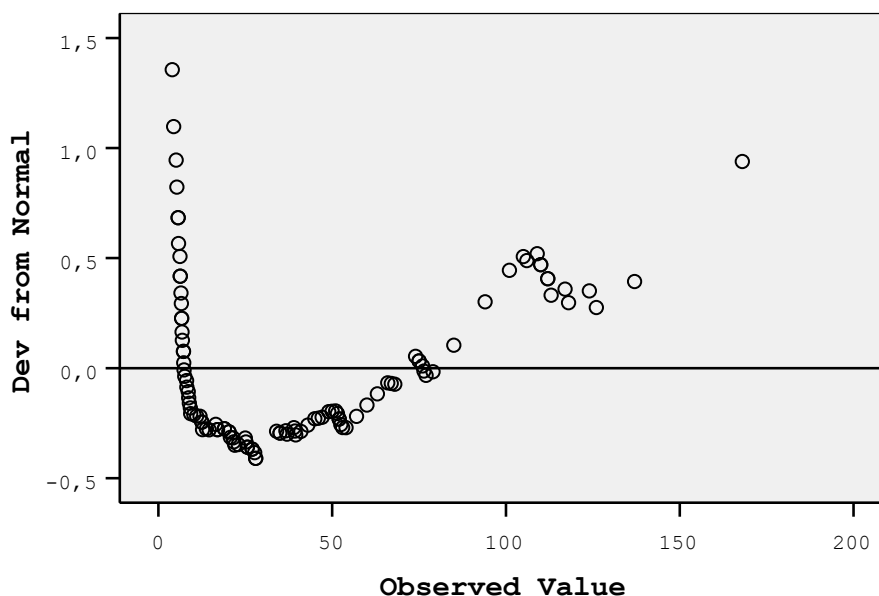
3.3.9 Graf pro ověřování normality dat (Q-Q plot)

Normal Q-Q Plot of Dětská úmrtnost (na tisíc živých porodů)



Obrázek 15 Graf pro ověřování normality dat (Q-Q plot), Normal

Detrended Normal Q-Q Plot of Dětská úmrtnost (na tisíc živých porodů)



Obrázek 16 Graf pro ověřování normality dat (Q-Q plot), Detrended

Graf pro ověřování normality dat (Q-Q plot) slouží k optickému posouzení, zda data pocházejí z normálního rozdělení. Graf zachycuje závislost mezi kvantily zkoumaného a normálního rozdělení. Souřadnice na ose x znázorňují pozorované hodnoty. Tyto hodnoty odpovídají v pozorovaném empirickém rozdělení kvantilům pro určité p . Na ose y se vynášejí hodnoty kvantilů standardizovaného normálního rozdělení pro stejné p . Čím blíže referenční přímce se potom body nacházejí, tím lepší je shoda s normálním rozdělením. Tvar grafu rovněž napoví, o jaký typ porušení normality se jedná. Jedním z nejčastějších případů je problém extrémních hodnot, kdy se odchylky od normality výrazně projeví především v okrajových částech grafu.

Pro větší názornost je tento graf v IBM SPSS Statistics zobrazen ještě jednou v podobě, kde je odečtena trendová složka, tj. referenční přímka je vodorovná.

3.3.10 Testy normality

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Dětská úmrtnost (na tisíc živých porodů)	.169	109	.000	.860	109	.000

a. Lilliefors Significance Correction

Tabulka 10 Test normality

Spolu s předchozím grafem se zobrazí také tabulka s výsledky dvou různých neparametrických testů normality (Kolmogorov-Smirnovův a Shapiro-Wilkův test). V obou případech obsahuje tabulka testovou statistiku, stupně volnosti a significance testu. Nulovou hypotézu normálního rozložení tedy zamítáme na 95% hladině spolehlivosti na základě daného testu v případě, že hodnota significance < 0.05 .



Jsou-li v datovém souboru zavedeny neceločíselné váhy, počítá se Shapiro-Wilkův test jen tehdy, je-li součet vah mezi 3 a 50. Pokud nejsou data vážena nebo při celočíselných vahách se statistika počítá, je-li součet vah mezi 3 a 5000.

Pozn.: Kolmogorov-Smirnovův test je v IBM SPSS Statistics k dispozici také mezi neparametrickými testy (NPAR TESTS). Různé výsledky těchto dvou procedur jsou způsobeny tím, že algoritmus procedury Explore vychází z tzv. Lillieforsovy korekce. Tato korekce je vhodná tam, kde známe pouze výběrový průměr a směrodatnou odchylku, zatímco procedura v neparametrických testech vychází z předpokladu o známých parametrech normálního rozdělení.

3.3.11 Grafy závislosti variability na poloze



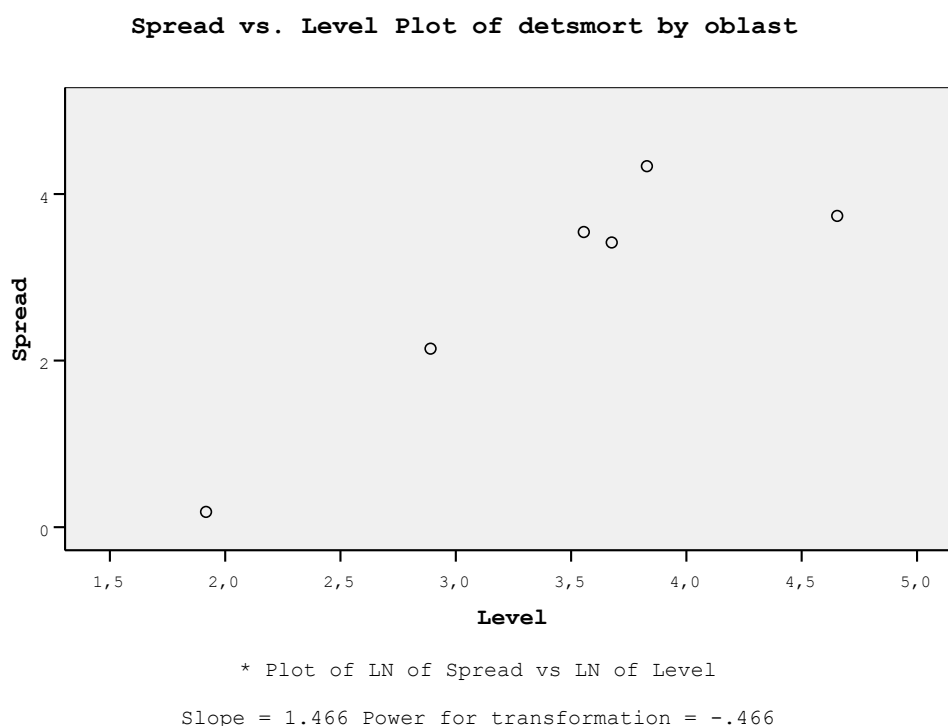
Obrázek 17 Grafy závislosti variability na poloze (1)

Grafy závislosti variability na poloze umožňují vizuálně posoudit homogenitu (podobnost) jednotlivých skupin (datových segmentů). Každý bod grafu charakterizuje jednu skupinu. Na horizontální osu x se vynáší hodnota mediánu ve skupině,

vertikální osa y reprezentuje kvartilové rozpětí. V případě, že byla požadována některá z transformací, zobrazují se hodnoty mediánu i kvartilového rozpětí transformované.

Tento typ grafů je určen především ke zjišťování závislosti heterogenity skupin na poloze a k ověření, zda jsou splněny předpoklady některých statistických testů (shoda rozptylů ve skupinách, homoskedasticita). Z grafu vyčteme, do jaké míry se liší variabilita jednotlivých skupin a vizuálně posoudíme, zda existuje vztah mezi variabilitou a polohou skupiny. Grafy transformací napoví, která z transformací by mohla nejlépe zajistit splnění daných předpokladů.

Hodnota *Slope* vyjadřuje směrnici (sklon) regresní přímky proložené body grafu.



Obrázek 18 Grafy závislosti variability na poloze (2)

V případě, že byla pro transformaci vybrána volba *Power*, zobrazí se v grafu ještě další koeficient *Power of transformation*, doporučující nejvhodnější exponent pro transformaci závislé proměnné typu obecná mocnina (tj. $Y = a \cdot X^p$). Tento koeficient se počítá pomocí iterativního algoritmu založeného na maximální věrohodnosti a vyjadřuje Box-Coxův parametr lambda, který doporučuje transformaci, která co nejlépe odstraní korelaci mezi variabilitou a polohou skupin. Potřeba takové transformace vychází z předpokladů některých metod a testů (například ANOVA). Je-li lambda = 1, není třeba žádná transformace, lambda = 0,5 odpovídá druhé odmocnině, při lambda = 0 je doporučen přirozený logaritmus, lambda = -0,5 charakterizuje převrácenou hodnotu druhé odmocniny a lambda = -1 převrácenou hodnotu.

V tomto případě je hodnota *Power of transformation* rovna -0.466, což odpovídá přibližně převrácené hodnotě druhé odmocniny.

3.3.12 Test shody rozptylů

Test of Homogeneity of Variance					
		Levene Statistic	df1	df2	Sig.
Dětská úmrtnost (na tisíc živých porodů)	Based on Mean	14.402	5	103	.000
	Based on Median	10.778	5	103	.000
	Based on Median and with adjusted df	10.778	5	46.756	.000
	Based on trimmed mean	13.051	5	103	.000

Tabulka 11 Test shody rozptylů

Předchozí graf je doplněn tabulkou s různými modifikacemi Levenova testu pro shodu rozptylů ve skupinách.





SEZNAM POUŽITÉ LITERATURY

Řehák, J., Brom, O. (2015). SPSS - Praktická analýza dat, Brno, Česká republika: Computer Press.



EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



SEZNAM OBRÁZKŮ

Obrázek 1: Nastavení dialogu Descriptives	6
Obrázek 2: Nastavení Descriptives, tlačítko Options	7
Obrázek 3: Nastavení dialogu Frequencies	10
Obrázek 4: Nastavení dialogu Frequencies, tlačítko Statistics.....	11
Obrázek 5: Nastavení dialogu Frequencies, tlačítko Charts.....	12
Obrázek 6: Nastavení dialogu Frequencies, tlačítko Format	13
Obrázek 7: Výstup procedury Frequencies, graf	15
Obrázek 8: Nastavení dialogu Explore	16
Obrázek 9: Nastavení dialogu Explore, tlačítko Statistics	17
Obrázek 10: Nastavení dialogu Explore, tlačítko Plots.....	18
Obrázek 11: Nastavení dialogu Explore, tlačítko Options	19
Obrázek 12 Boxplot.....	23
Obrázek 13 Histogram.....	24
Obrázek 14 Stem&Leaf graf	25
Obrázek 15 Graf pro ověřování normality dat (Q-Q plot), Normal	26
Obrázek 16 Graf pro ověřování normality dat (Q-Q plot), Detrended	26
Obrázek 17 Grafy závislosti variability na poloze (1).....	28
Obrázek 18 Grafy závislosti variability na poloze (2).....	29





SEZNAM TABULEK

Tabulka 1 Výstup procedury Descriptives	8
Tabulka 2 Výstup procedury Descriptives - seřazeno	9
Tabulka 3 Výstup procedury Frequencies, Statistics	14
Tabulka 4 Výstup procedury Frequencies, tabulka četností	15
Tabulka 5 Přehled o počtu platných a chybějících případů	20
Tabulka 6 Popisné statistiky z procedury Explore	20
Tabulka 7 Odhady střední hodnoty	21
Tabulka 8 Percentily	22
Tabulka 9 Extreme Values	22
Tabulka 10 Test normality	27
Tabulka 11 Test shody rozptylů	30





EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY