



## Explore – podrobná analýza dat

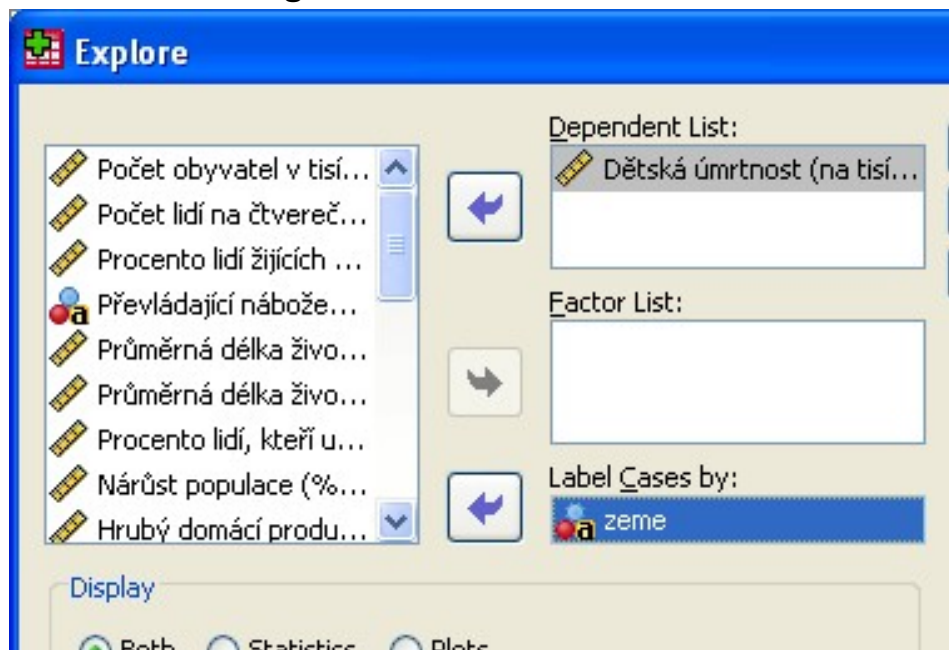
### K čemu slouží procedura Explore

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.

### Volání procedury v IBM SPSS Statistics

Analyze → Descriptive Statistics → Explore

### Nastavení dialogu

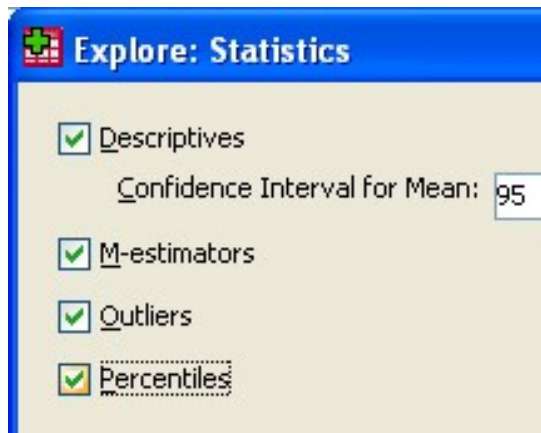


- Do okna *Dependent List* přeneseme analyzované číselné proměnné.
- V poli *Factor List* zadáváme jednu nebo více kategorizovaných proměnných, které rozdělí datový soubor na skupiny. Tyto skupiny soubor v analýze člení a jsou buď samostatně a/nebo komparačně (graficky) popisovány. V případě, že zadáme více faktorů, provedou se analýzy pro skupiny určené každým z nich zvlášť.
- Do okna *Label Cases by* umísťujeme proměnnou, která určuje názvy jednotlivých případů.

## Listy procedur IBM SPSS Statistics

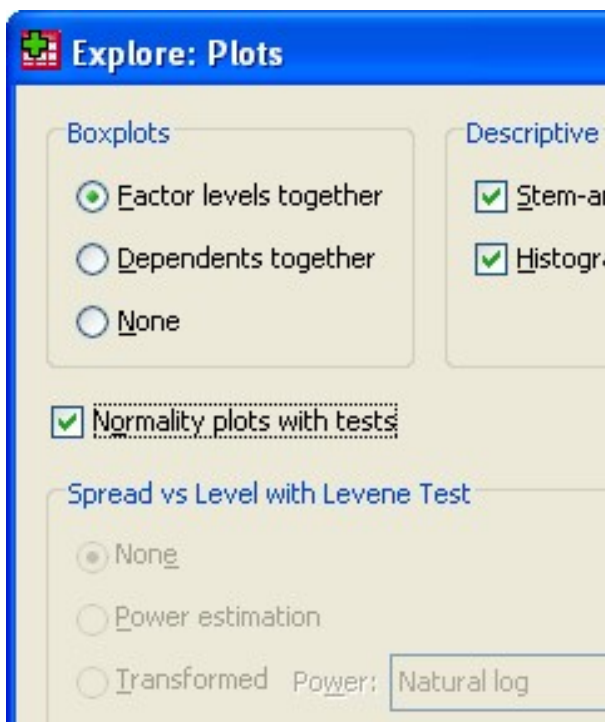
- V části okna označené *Display* rozhodujeme, zda nás zajímají pouze statistiky, pouze grafy nebo obojí.

### 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.

## Tlačítko Plots

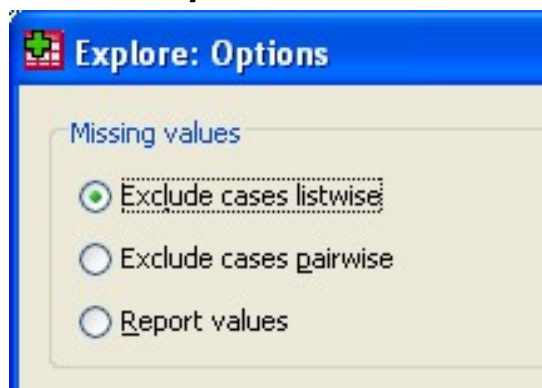


- Přepínač *Boxplots* určuje zda, a případně jakým způsobem, se vytvoří boxplot. Pro každou zkoumanou proměnnou lze nakreslit samostatný diagram (*Factor levels together*), ve kterém budou grafy pro jednotlivé skupiny, nebo můžeme vytvořit jediný graf pro porovnávání vlastností všech analyzovaných proměnných (*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 heterogenity 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*.

## Tlačítko Options



Přepínačem *Missing Values* určujeme, 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*).

## Výstupy

### 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 poskytuje informaci o počtu platných a chybějících hodnot u analyzovaných proměnných.

### 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 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.

## Odhady střední hodnoty

**M-Estimators**

	Huber's M-Estimator <sup>a</sup>	Tukey's Biweight <sup>b</sup>	Hampel's M-Estimator <sup>c</sup>	Andrews' Wave <sup>d</sup>
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  $1.340 \cdot \pi$ .

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. Extrémní 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.

## Percentily

**Percentiles**

		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		

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).

## Výpis pěti nejnižších a nejvyšších hodnot

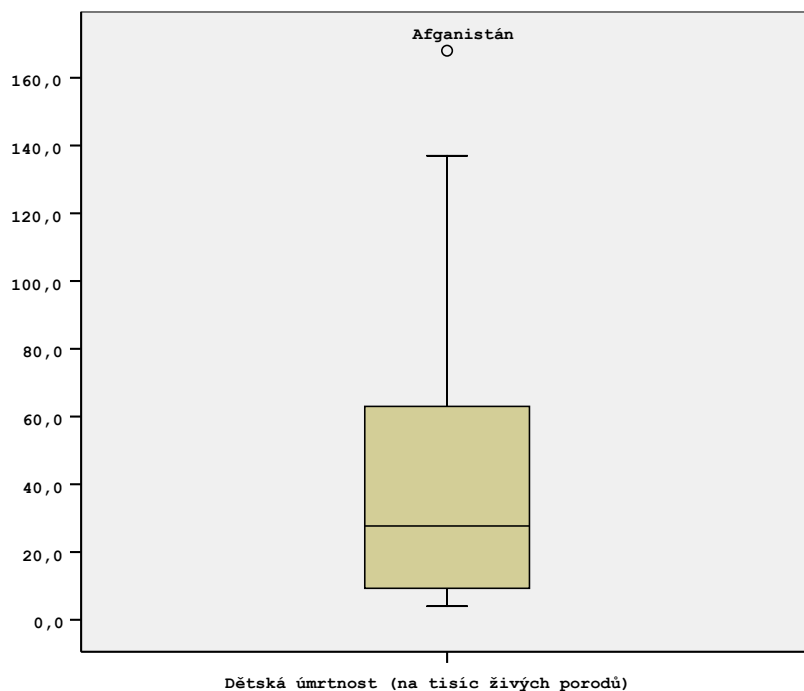
Extreme Values					
			Case Number	zeme	Value
Dětská úmrtnost (na tisíc živých porodů)	Highest	1	1	Afganistán	168.0
		2	22	JAR	137.0
		3	90	Somálsko	126.0
		4	40	Gambie	124.0
		5	17	Burkina Faso	118.0
	Lowest	1	49	Island	4.0
		2	57	Japonsko	4.4
		3	96	Taiwan	5.1
		4	37	Finsko	5.3
		5	93	Švédsko	5.7 <sup>a</sup>

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

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.

## Boxplot

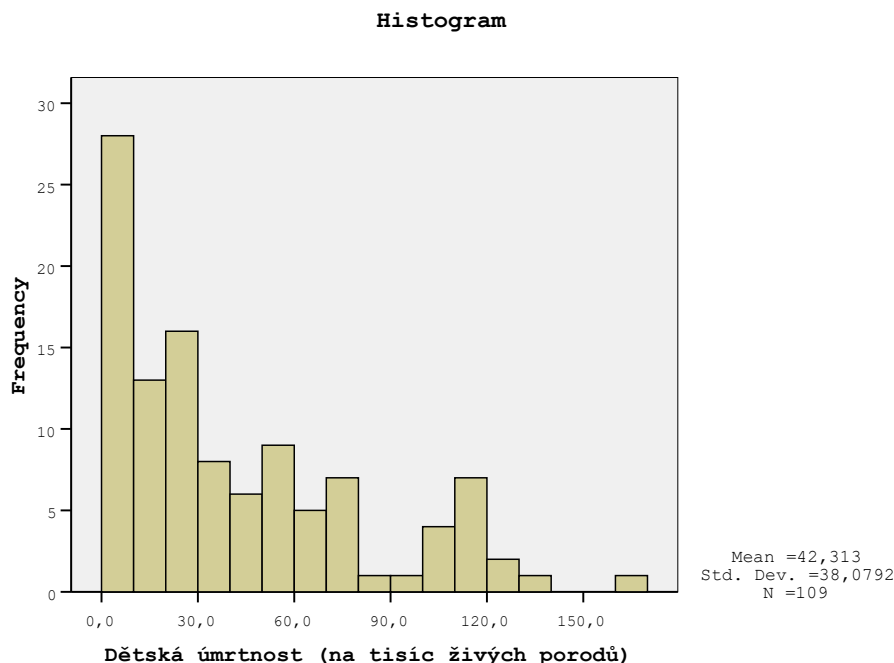


Boxplot (krabicový graf) poskytuje základní informaci o rozložení proměnné a názorně zachycuje medián, kvartily i extrémy v datech. Dolní a horní okraj krabičky odpovídají kvartilům, dělící čára uvnitř krabičky vyjadřuje medián. Jestliže se všechny případy nacházejí maximálně ve vzdálenosti 1,5-násobku 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

## Listy procedur IBM SPSS Statistics

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 extrémní hodnoty, které překročily hranici 3-násobku kvartilového rozpětí (v grafu se značí hvězdičkou).

### 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.



### Listový (cifrový) diagram

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

Frequency	Stem &	Leaf
28.00	0 .	44555556666666666777778888899
13.00	1 .	0122223467799
16.00	2 .	000112355557788
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)

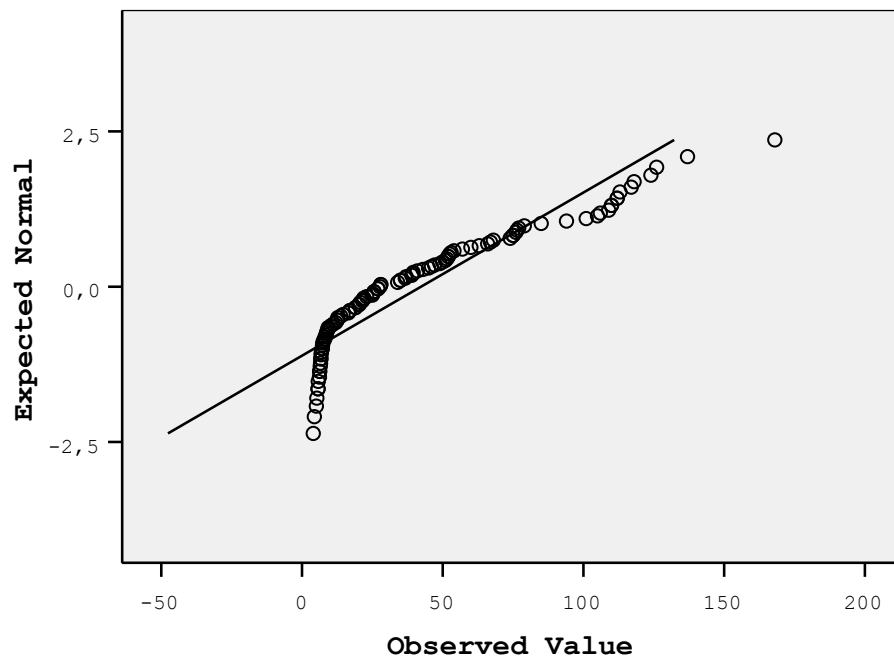
Stem width: 10.0  
Each leaf: 1 case(s)

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.

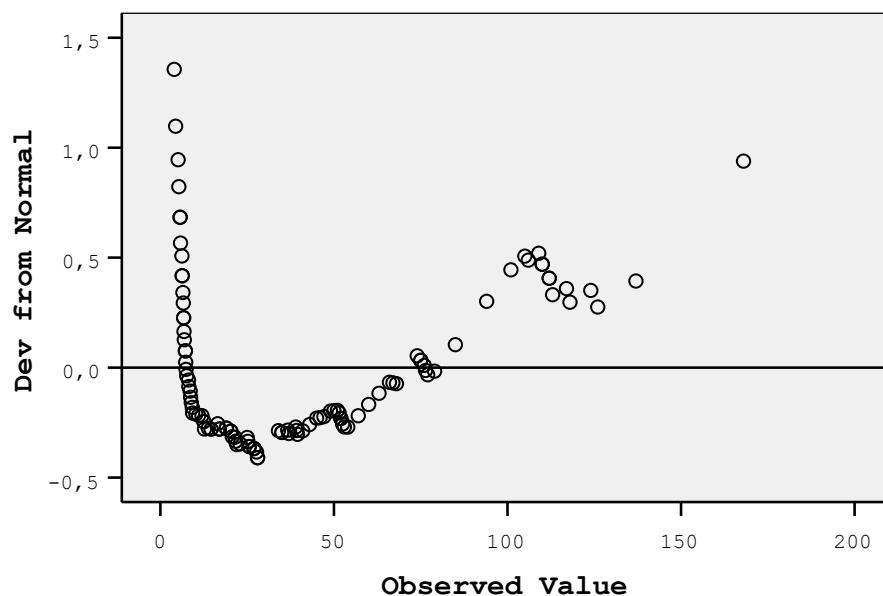
Výhodou cifrového diagramu proti histogramu je větší množství obsažené informace – data jsou seříděna a seskupena a známe nejen četnosti intervalů, ale také všechny originální hodnoty. Graf lze rovněž velmi snadno vytvářet i ručně a průběžně doplňovat další hodnoty tak, jak jsou sbírány, což bylo dříve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

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

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



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



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

## Listy procedur IBM SPSS Statistics

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.

### Testy normality

Tests of Normality						
	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Dětská úmrtnost (na tisíc živých porodů)	.169	109	.000	.860	109	.000

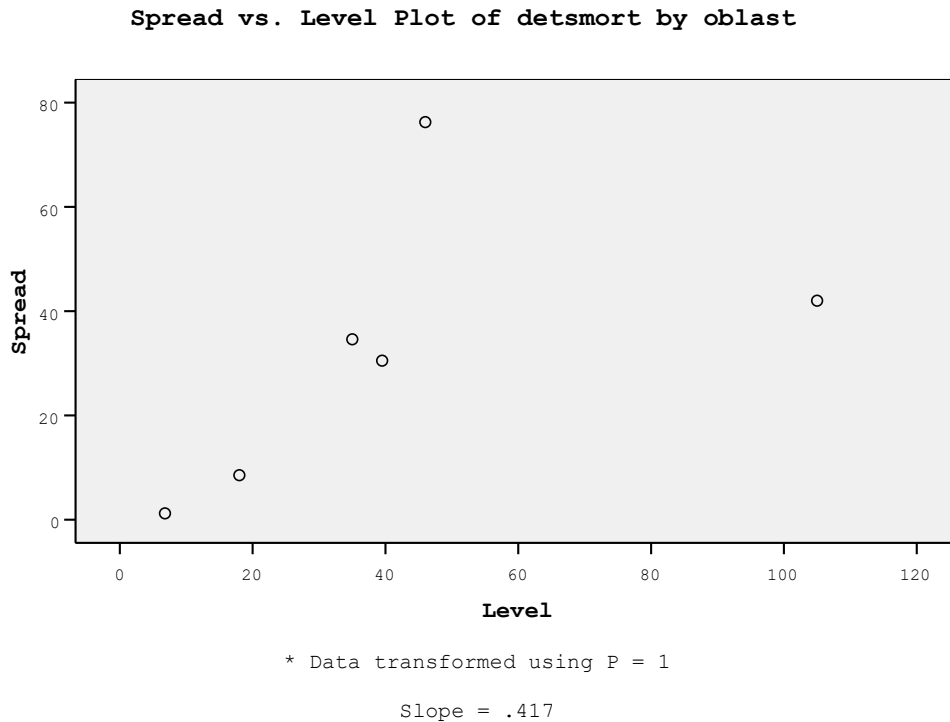
a. Lilliefors Significance Correction

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.

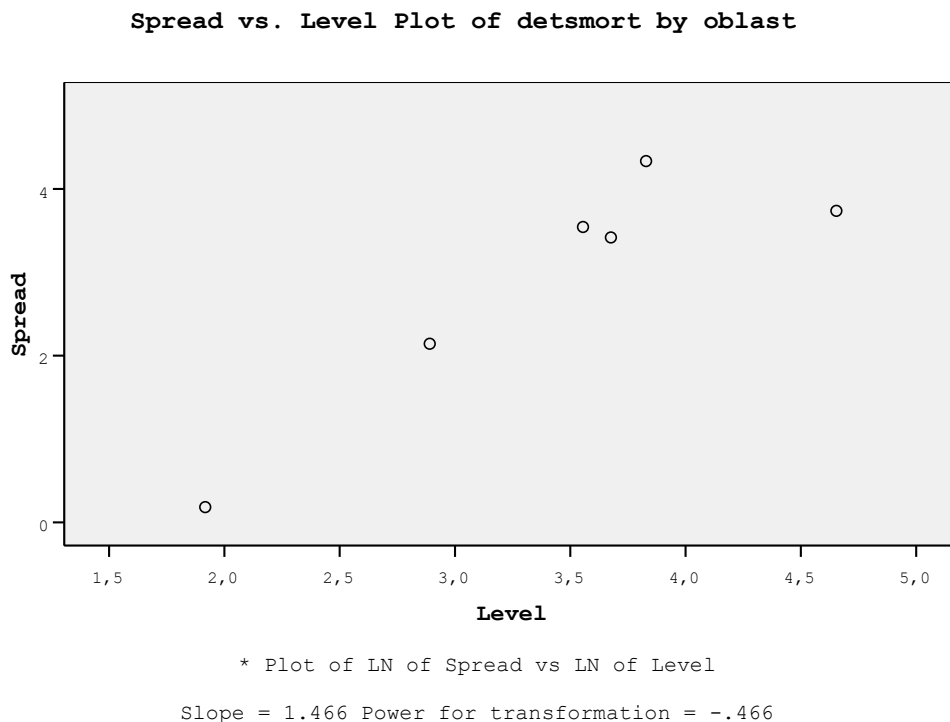
### Grafy závislosti variability na poloze



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.



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.

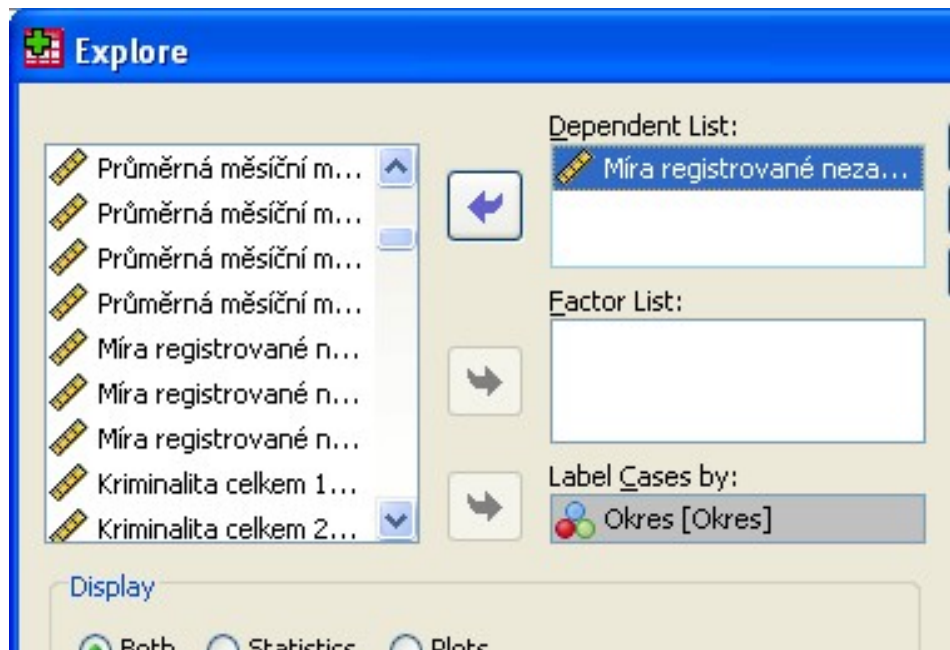
## Test shody rozptylů

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

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

### Příklad

Datový soubor obsahuje údaje o okresech České republiky, každý řádek představuje jeden okres. Naším cílem je získat základní informace o míře nezaměstnanosti a jejím rozdělení a současně provést srovnání krajů. Data o nezaměstnanosti jsou k dispozici za rok 2000 – 2003 a jsou uložena ve 4 samostatných proměnných odpovídajícím jednotlivým rokům.



Nejprve budeme analyzovat samostatně proměnnou *Míra registrované nezaměstnanosti 2003*. V zadání procedury *Explore* přeneseme tuto proměnnou do pole *Dependent List*. Vzhledem k tomu, že u některých výstupů je třeba přesně vědět, o který okres se jedná, v poli *Label Cases by* zadáme jako popis případů proměnnou vyjadřující názvy okresů.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Míra registrované nezaměstnanosti 2003	77	100.0%	0	.0%	77	100.0%

Následující tabulka nás informuje, že v našem datovém souboru je celkem 77 okresů (Hlavní město Praha je spojeno a představuje jediný případ) a u všech je uvedena hodnota proměnné *Míra registrované nezaměstnanosti 2003*.

## Listy procedur IBM SPSS Statistics

**Descriptives**

			Statistic	Std. Error
Míra registrované nezaměstnanosti 2003	Mean		10.465	.4855
	95% Confidence Interval for Mean	Lower Bound	9.498	
		Upper Bound	11.433	
	5% Trimmed Mean		10.263	
	Median		9.695	
	Variance		18.153	
	Std. Deviation		4.2606	
	Minimum		3.0	
	Maximum		23.5	
	Range		20.5	
	Interquartile Range		5.1	
	Skewness		.862	.274
	Kurtosis		.503	.541

Z tabulky popisných statistik zjistíme, že minimální hodnota nezaměstnanosti v roce 2003 představovala 3 %, maximální hodnota 23,5 %, rozdíl mezi okresem s nejvyšší a nejnižší nezaměstnaností je celých 20,5 %. Průměrná hodnota (10,5 %) je o něco vyšší než useknutý průměr (10,3 %) i medián (9,7 %), což naznačuje, že rozdělení bude pravděpodobně zešíkmené, a problém mohou představovat extrémně vysoké hodnoty.

**M-Estimators**

	Huber's M-Estimator <sup>a</sup>	Tukey's Biweight <sup>b</sup>	Hampel's M-Estimator <sup>c</sup>	Andrews' Wave <sup>d</sup>
Míra registrované nezaměstnanosti 2003	9.826	9.462	9.882	9.445

- 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\*pi.

Rovněž nižší hodnoty v tabulce odhadů střední hodnoty v porovnání s průměrem napovídají o tom, že průměr je ovlivněn extrémny s vysokými hodnotami nezaměstnanosti.

**Percentiles**

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average(Definition 1)	Míra registrované nezaměstnanosti 2003	4.603	5.487	7.689	9.695	12.825	18.174	18.988
Tukey's Hinges	Míra registrované nezaměstnanosti 2003			7.729	9.695	12.504		

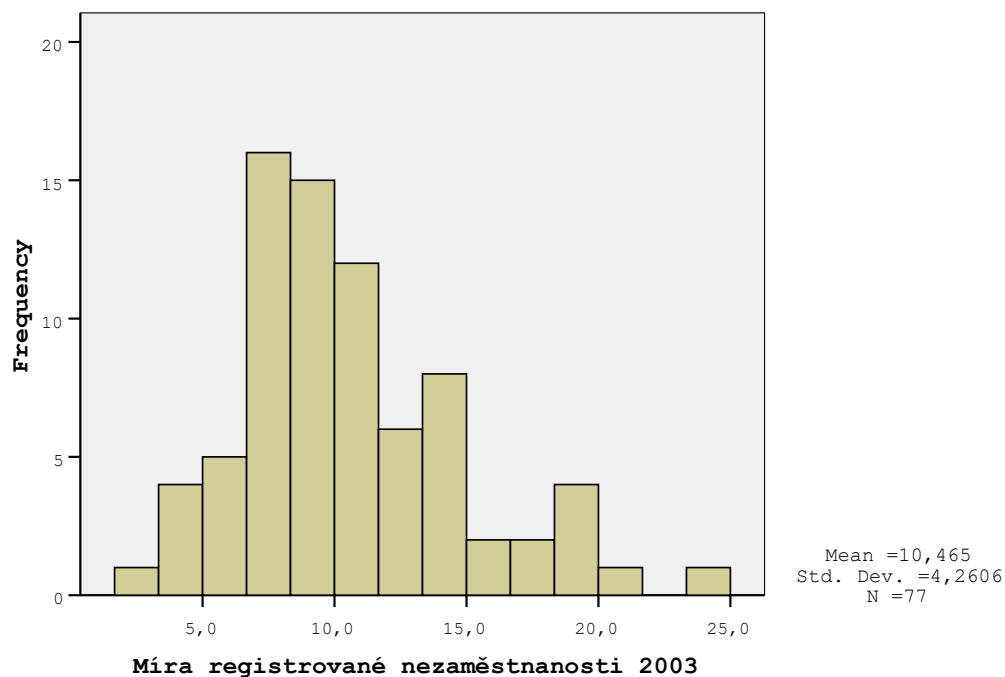
Z tabulky percentilů vyčteme, že 5 % okresů má nezaměstnanost do 4,6 %, 10 % okresů má nezaměstnanost do 5,5 %, ale naopak 5 % okresů má nezaměstnanost 19 % nebo více.

Extreme Values

			Case Number	Okres	Value
Míra registrované nezaměstnanosti 2003	Highest	1	35	Most	23.5
		2	74	Karviná	20.4
		3	36	Teplice	19.9
		4	34	Louny	18.9
		5	32	Chomutov	18.7
	Lowest	1	11	Praha-západ	3.0
		2	10	Praha-východ	3.8
		3	1	Praha	4.0
		4	2	Benešov	4.7
		5	14	České Budějovice	4.8

Mezi okresy s nejvyšší nezaměstnaností patří Most, Karviná, Teplice, Louny a Chomutov. Naopak nejnižší hodnoty byly zjištěny v Praze, Benešově a Českých Budějovicích.

Histogram



Histogram potvrzuje podezření o zešíkmeném rozdělení. Samostatný sloupeček v pravé části grafu vypovídá o tom, že se v datech vyskytují některé okresy s dost vysokou mírou nezaměstnanosti.



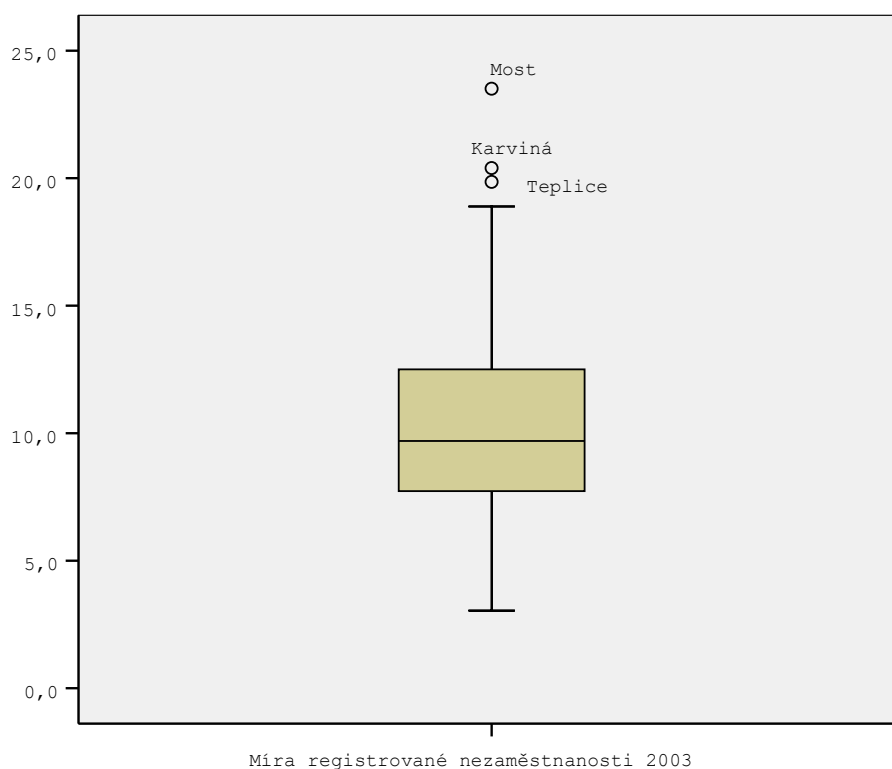
## Listy procedur IBM SPSS Statistics

### Míra registrované nezaměstnanosti 2003 Stem-and-Leaf Plot

Frequency	Stem &	Leaf
2.00	0 .	33
7.00	0 .	4445555
12.00	0 .	666667777777
20.00	0 .	8888888888888999999
14.00	1 .	00000000001111
7.00	1 .	2223333
7.00	1 .	4444455
.00	1 .	
5.00	1 .	88888
3.00	Extremes	(>=20)

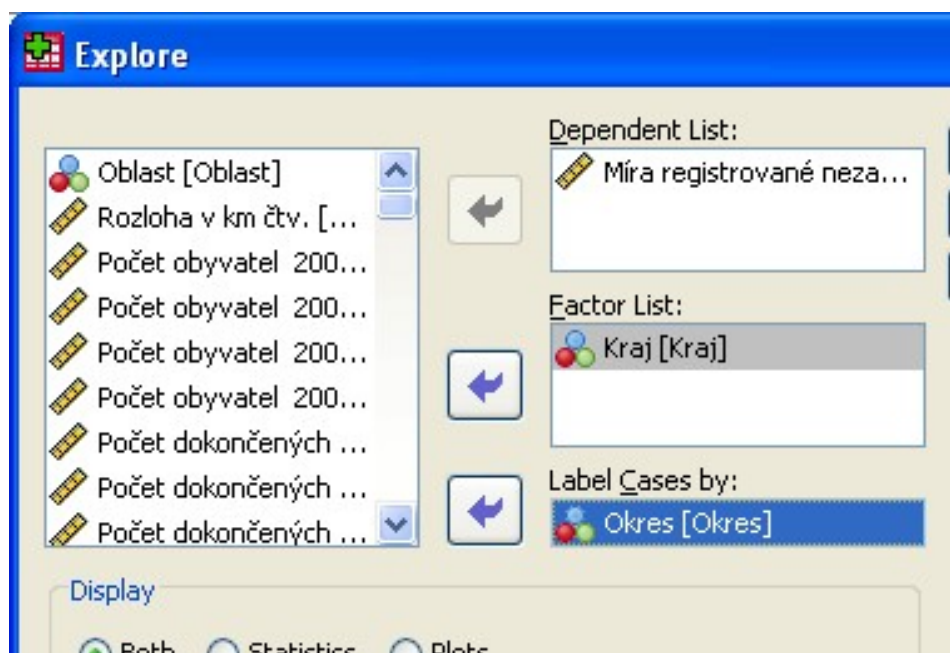
Stem width: 10.0  
Each leaf: 1 case(s)

Z cifrového grafu je ještě lépe vidět problém vysokých hodnot, které mohou při některých typech analýz způsobovat problémy. Celkem ve 3 okresech dosahují hodnoty nezaměstnanosti více než 20 %.

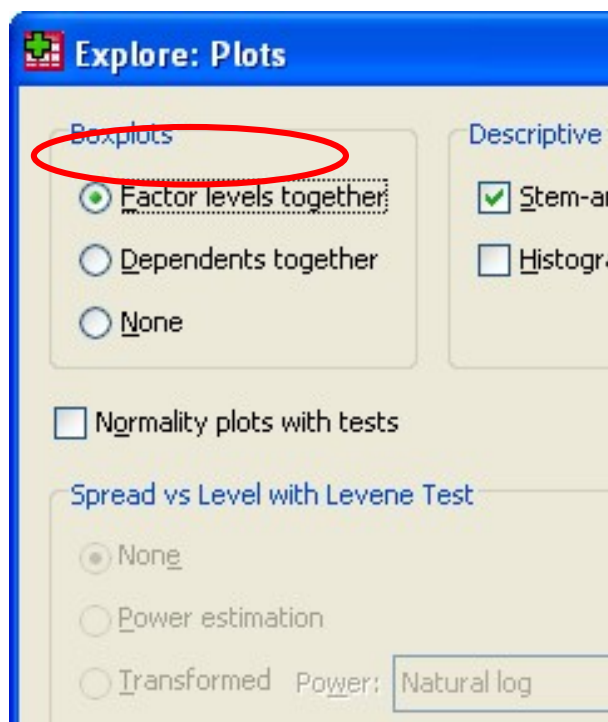


Z boxplotu zjistíme, že tři okresy – Most, Karviná a Teplice – byly klasifikovány jako odlehlá pozorování, což znamená, že se chovají poněkud jinak než zbytek souboru. Větší protažení grafu v horní části opět vyjadřuje šikmost rozdělení.

## Listy procedur IBM SPSS Statistics



Nyní provedeme jednoduché srovnání nezaměstnanosti v krajích. V dialogovém okně procedury *Explore* zůstane stejné zadání, pouze do pole *Factor List* přidáme proměnnou *Kraj*.



Pomocí tlačítka *Plots* vybereme v části *Boxplots* volbu *Factor levels together*, která umožní grafické porovnání krajů.

## Listy procedur IBM SPSS Statistics

### Descriptives<sup>a</sup>

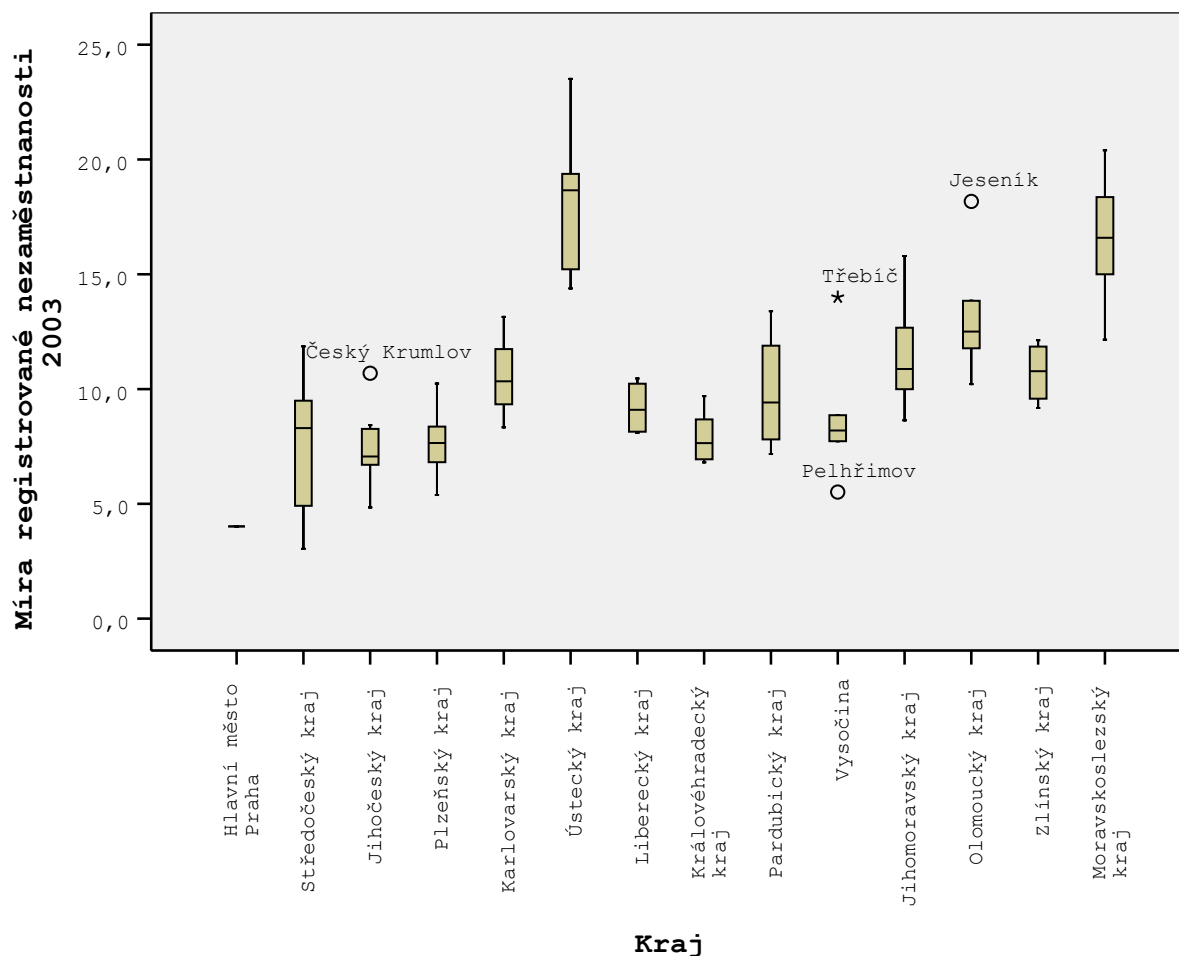
Míra registrované nezaměstnanosti 2003

Statistic

Kraj	Mean	95% Confidence Interval for Mean		5% Trimmed Mean	Median	Variance	Std. Deviation	Minimum	Maximum	Range	Interquartile Range	Skewness	Kurtosis
		Lower Bound	Upper Bound										
Středočeský kraj	7.40	5.59	9.21	7.39	8.30	8.13	2.85	3.04	11.86	8.82	4.89	-.15	-1.31
Jihočeský kraj	7.50	5.81	9.20	7.47	7.06	3.35	1.83	4.84	10.69	5.84	2.04	.49	1.06
Plzeňský kraj	7.66	6.23	9.09	7.64	7.65	2.38	1.54	5.38	10.24	4.86	1.66	.30	.59
Královéhradecký kraj	7.95	6.43	9.47	7.92	7.65	1.50	1.22	6.81	9.69	2.88	2.31	.70	-1.17
Vysočina	8.86	4.97	12.75	8.76	8.19	9.84	3.14	5.51	14.00	8.49	4.81	1.31	2.67
Liberecký kraj	9.19	7.24	11.13	9.18	9.10	1.49	1.22	8.09	10.46	2.37	2.23	.12	-5.28
Pardubický kraj	9.85	5.54	14.15	9.80	9.42	7.32	2.71	7.17	13.39	6.22	5.15	.76	-.29
Karlovarský kraj	10.61	4.60	16.61	.	10.34	5.84	2.42	8.34	13.14	4.81	.	.49	.
Zlínský kraj	10.71	8.54	12.89	10.72	10.78	1.88	1.37	9.17	12.13	2.95	2.61	-.15	-3.62
Jihomoravský kraj	11.52	9.23	13.81	11.44	10.87	6.14	2.48	8.64	15.80	7.16	4.17	.84	.07
Olomoucký kraj	13.30	9.55	17.05	13.20	12.50	9.12	3.02	10.22	18.17	7.95	5.01	1.24	1.89
Moravskoslezský kraj	16.51	13.37	19.66	16.54	16.59	8.99	3.00	12.15	20.40	8.25	4.58	-.23	-.87
Ústecký kraj	17.96	14.94	20.98	17.85	18.66	10.64	3.26	14.38	23.51	9.13	5.01	.60	-.25

a. Míra registrované nezaměstnanosti 2003 is constant when Kraj = Hlavní město Praha. It has been omitted.

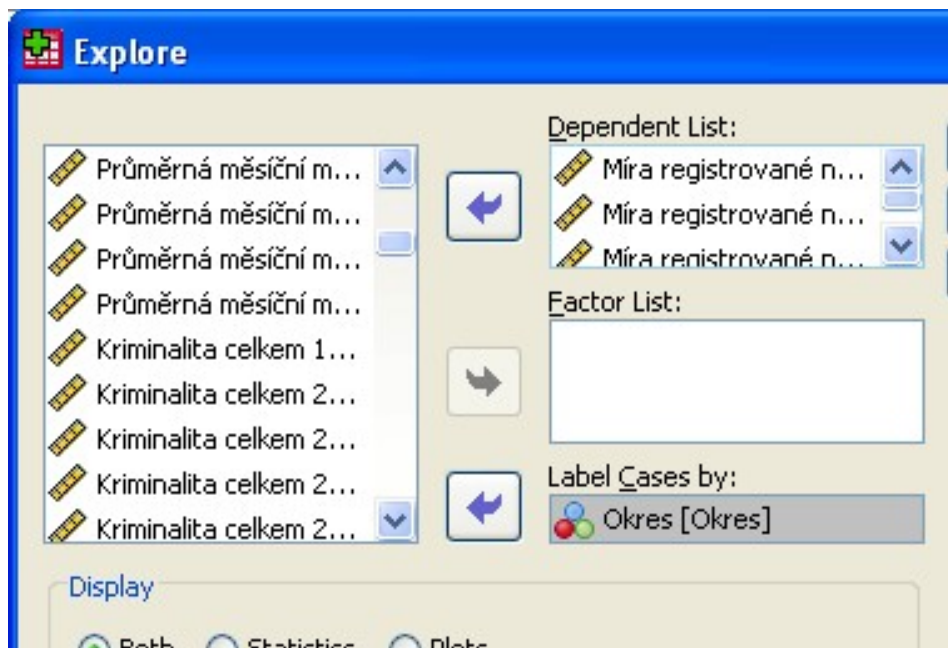
Tabulka popisných statistik byla z důvodu přehlednosti upravena pomocí pivotace tak, aby v řádcích byly umístěny kraje, ve sloupcích popisné statistiky. Název analyzované proměnné a přepínání mezi statistikou a její standardní chybou se nacházejí ve vrstvách. Okresy jsou rovněž seřazeny podle průměrné míry nezaměstnanosti. Tabulka poskytuje základní srovnání popisných charakteristik za kraje. Vzhledem k tomu, že kraj Praha je reprezentován pouze jednou hodnotou, statistiky se zde nezobrazují.



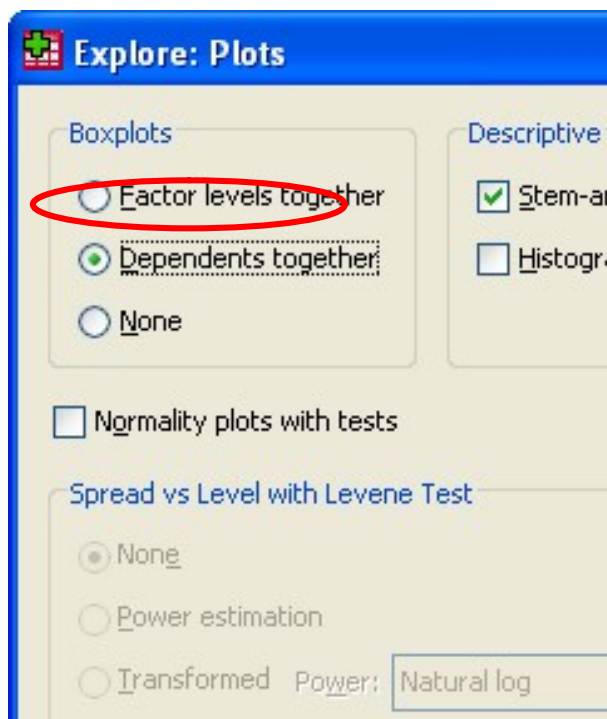
Boxplot zcela jasně ukazuje, že největší problémy s nezaměstnaností jsou v Ústeckém a Moravskoslezském kraji. Mezi kraje s vysokou nezaměstnaností patří rovněž Olomoucký kraj, kde se nejvýrazněji projevuje tato tendence v okrese Jeseník. Vzhledem k tomu, že Praha v datech není rozdělena na okresy a představuje pouze jeden řádek, reprezentuje graf pouze úsečka v uvedené hodnotě. Pokud provedeme srovnání jenom na základě mediánů, vychází kromě Prahy nejlépe Jihočeský a Plzeňský kraj. Ve Středočeském kraji pozorujeme značnou variabilitu mezi okresy, naopak nezaměstnanost v okresech Libereckého nebo Královéhradeckého kraje je velmi vyrovnaná.

## Listy procedur IBM SPSS Statistics

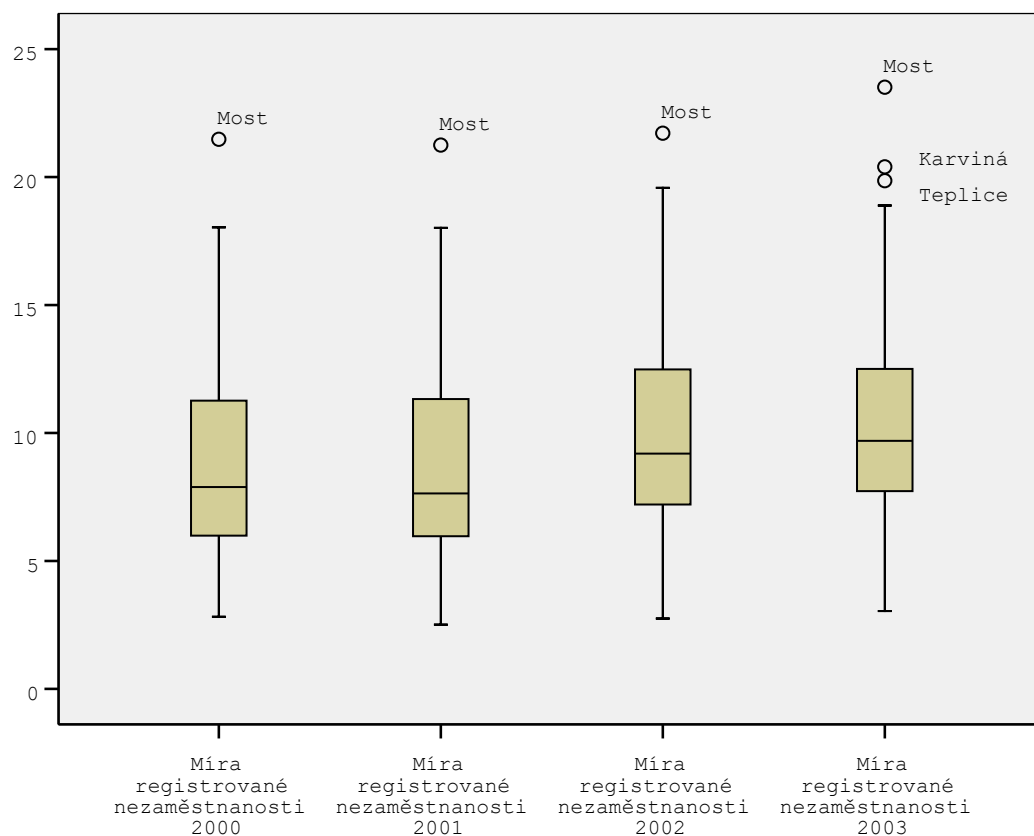
V následujícím kroku se pokusíme zjistit, jak se měnila nezaměstnanost v letech 2000 – 2003.



Do pole *Dependent List* přeneseme proměnné *Míra registrované nezaměstnanosti 2000 ... Míra registrované nezaměstnanosti 2003*, v poli *Label Cases by* zůstane popis okresů.



Pomocí tlačítka *Plots* označíme v části *Boxplots* volbu *Dependents together*, která umožní srovnání nezaměstnanosti v jednotlivých letech.



Z grafu vyplývá, že míra nezaměstnanosti se v průběhu čtyř let výrazně nemění, mírně se však zvyšuje. Ve všech případech se objevuje okres Most jako odlehlé pozorování s nejvyšší mírou nezaměstnanosti. V roce 2003 se zobrazují jako odlehlá pozorování rovněž okresy Karviná a Teplice, což je způsobeno jednak vyšší nezaměstnaností v těchto okresech, jednak o něco menším kvartilovým rozpětím v porovnání s předchozími roky.

## Listy procedur IBM SPSS Statistics

### Descriptives

Statistic		Míra registrované nezaměstnanosti 2000	Míra registrované nezaměstnanosti 2001	Míra registrované nezaměstnanosti 2002	Míra registrované nezaměstnanosti 2003
Mean		8.889	8.995	9.939	10.465
95% Confidence Interval for Mean	Lower Bound	7.960	8.090	8.997	9.498
	Upper Bound	9.819	9.901	10.881	11.433
5% Trimmed Mean		8.678	8.806	9.764	10.263
Median		7.886	7.637	9.195	9.695
Variance		16.765	15.921	17.221	18.153
Std. Deviation		4.0945	3.9901	4.1499	4.2606
Minimum		2.8	2.5	2.8	3.0
Maximum		21.5	21.3	21.7	23.5
Range		18.7	18.7	19.0	20.5
Interquartile Range		5.4	5.4	5.5	5.1
Skewness		.875	.810	.727	.862
Kurtosis		.297	.248	.051	.503

Tabulka popisných statistik je upravena pomocí pivotace tak, aby proměnné byly vedle sebe ve sloupcích. To umožňuje snadno porovnat jednotlivé statistiky v průběhu čtyř let. Rovněž z této tabulky vyplývá, že nezaměstnanost se mírně zvýšila, což je vidět především na hodnotách průměru a mediánu.