Deskriptivní statistika

POLb1139 Statistické myšlení v sociálních vědách

Dnes se posouváme o krok dále

- Známe typy proměnných
- Máme data (vlastní sběr / jiným způsobem)
  - 2 příklady: průzkum v předmětu (n=40) a ESS (n=2000)

- Jak začít analýzu?

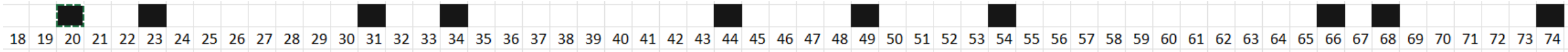
- Ideální první kroky:
  - Poznejte svá data – struktura, distribuce
  - Vizualizace dat

Deskriptivní analýza

- Explorace dat v rámci jedné proměnné
- Cílem je popsat a porozumět datům
- Má smysl před vícerozměrnou analýzou (nebo i samostatně)

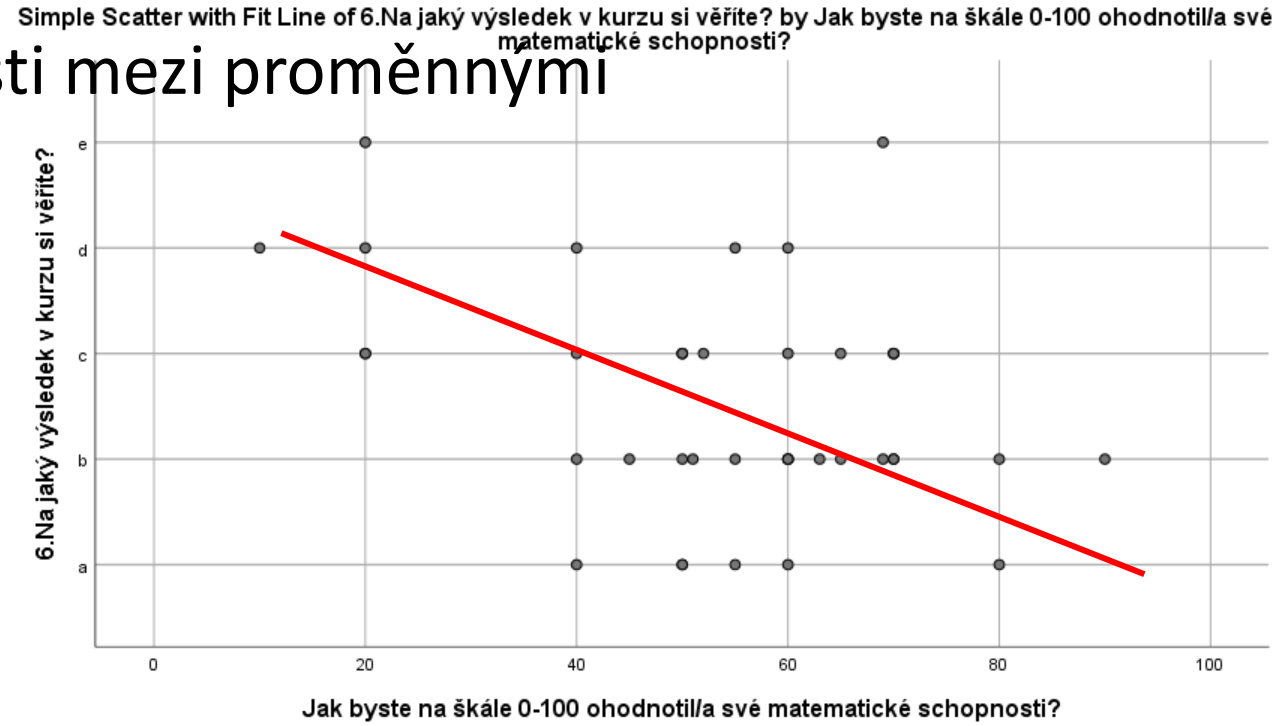
Jednorozměrná analýza

- Získané údaj se týkají vždy jen dané proměnné



- Nehledáme rozdíly ani souvislosti mezi proměnnými

- Dvourozměrná analýza
  - Souvislost mezi proměnnými
    - Kontingenční tabulky
    - Srovnání průměrů
    - korelace



Deskriptivní analýza

- Záleží na úrovni proměnných podle měření
  - Různé typy proměnných poskytují různé možnosti
  - Kardinální > ordinální > nominální
  - SPSS vás zpravidla nezachrání (a neupozorní na očividný nesmysl)

- Prostor pro odhalování chyb (měření)
  - Minimum a maximum
  - Identifikace odlehlých případů (outliers)
  - Identifikace chyb při vkládaní dat (pokud se dají jednoduše rozpoznat)

Nominální proměnné

- Pojmenování kategorie
- Co je (a není) s nimi možné dělat?
- kolik případů spadá do jednotlivých kategorií?
  - Četnost (anglicky frequency)
  - Modus (nejčastější četnost)
- Číselné kódy pro jejich hodnoty mají pouze symbolický význam → důsledky?

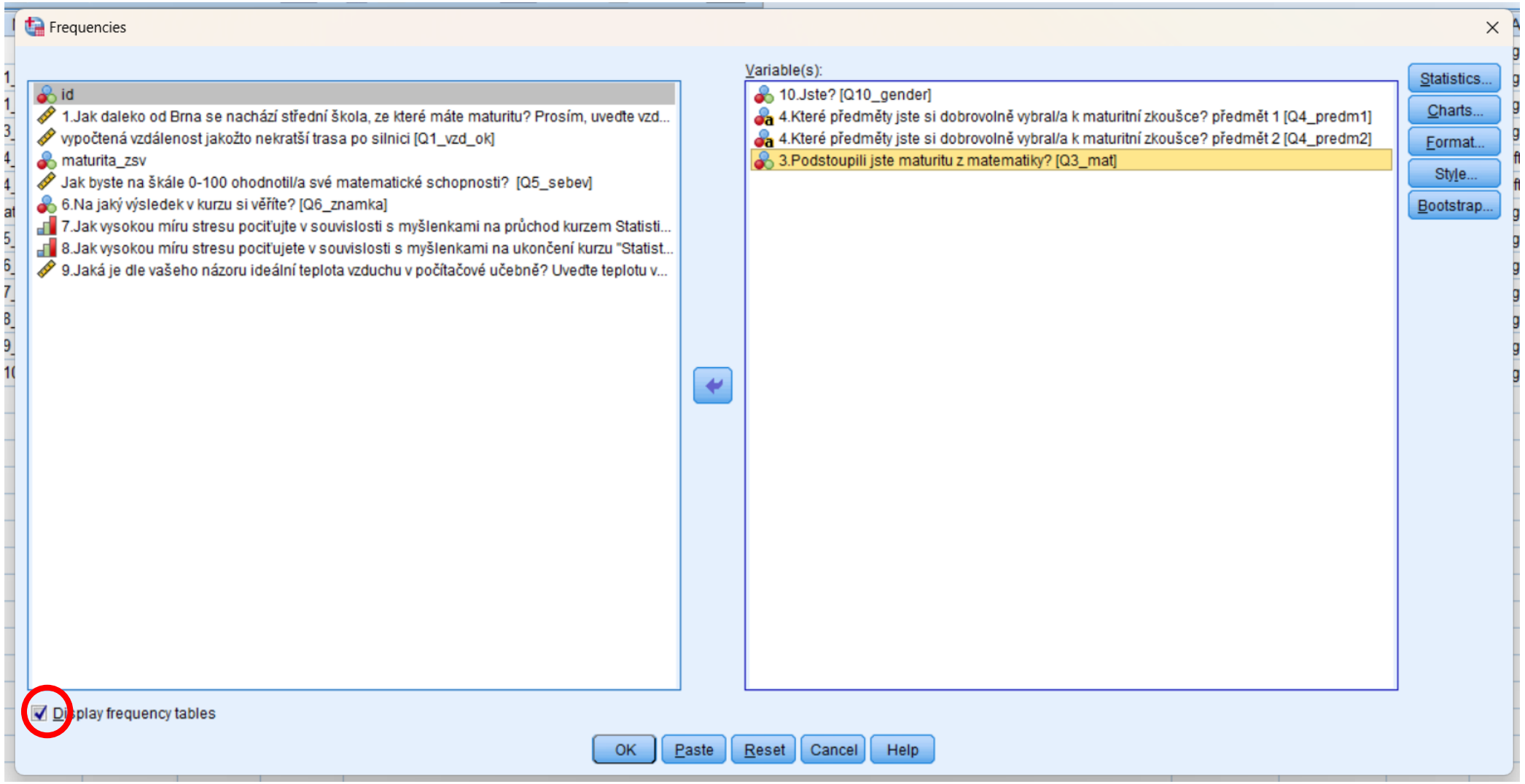
Modus

- Nejčastější hodnota
- Frekvenční tabulka - nejvyšší hodnota

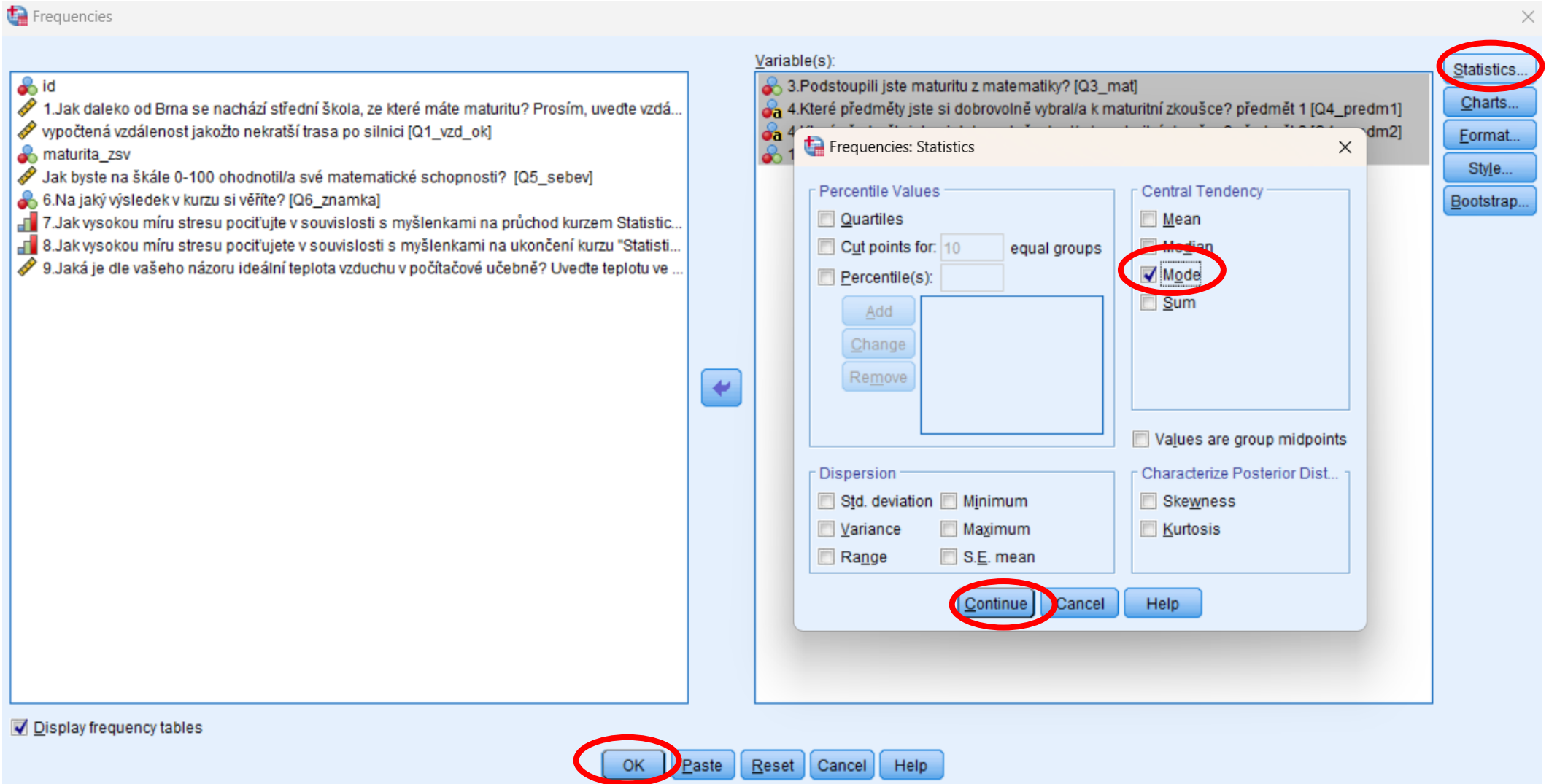
- Využití pro všechny typy proměnných
- Modus nemusí být nutně pouze jeden (bimodální, multimodální distribuce)

Tabulka četností

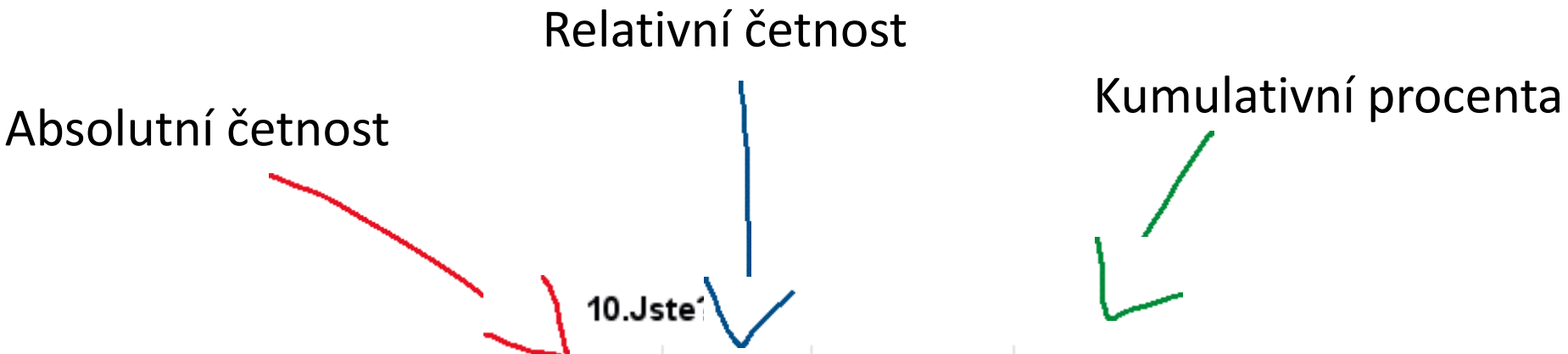
- Analyze → Descriptive Statistics → Frequencies



- Analyze → Descriptive Statistics → Frequencies



Tabulka četností



10.Jste

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	žena	16	40,0	43,2	43,2
	muž	21	52,5	56,8	100,0
	Total	37	92,5	100,0	
Missing	nechcei odpovědět	3	7,5		
Total		40	100,0		

Statistics

		3.Podstoupili jste maturitu z matematiky?	4.Které předměty jste si dobrovolně vybral/a k maturitní zkoušce? předmět 2	10.Jste?
N	Valid	40	40	37
	Missing	0	0	3
Mode		0		2

Jaký je modus proměnné Q4_predm1?

Ordinální proměnné

- Lze seřadit

- Četnosti
- Modus
- Medián

Medián

- Středová hodnota, rozděluje dataset na dvě poloviny hodnot
  - Hodnota, pod kterou leží 50 % (polovina) hodnot a nad kterou leží 50 % (polovina) hodnot
  - V ordinálních datech = mediánová kategorie (kumulativní četnost zahrnuje 50 % případů pod mediánem)
  - 50. percentil
  - "*My dnes víme, že 50 procent obyvatel má mzdu pod úrovní mediánu*." Jiří Šlégr (2016)

- Postup:
  - Seřadíme hodnoty vzestupně
  - Najdeme tu, která leží uprostřed data setu (jednodušší pro matice s lichým počtem hodnot)

- Výhoda: je stabilní, není citlivý na extrémní hodnoty

Medián - příklad

- Počet hodin denně na sociálních sítích (9 lidí): 7, 0, 15, 8, 4, 6, 3, 10, 1
  - Seřazení → 0, 1, 3, 4, **6**, 7, 8, 10, 15
  - Výběr hodnoty uprostřed (5. v pořadí) → **6**
  - **4 lidé mají nižší hodnotu, 4 vyšší**

- Co když máme sudý počet pozorování?
  - 8 lidí, stejný příklad: 7, 0, 15, 8, 4, 6, 3, 10
  - Seřazení → 0, 3, 4, **6, 7**, 8, 10, 15
  - Medián je uprostřed dvou prostředních naměřených hodnot: (6+7)/2 = **6,5**

- Sudý a lichý počet – při velkém počtu dat je rozdíl věcně zanedbatelný

	id	Q1_vzd	Q1_vzd_ok	Q3_mat	Q4_predm1	Q4_predm2	maturita_zsv	Q5_sebev	Q6_znamka
1	22	268	311	1	matematika		ne	60	a
2	7	92	90	1	zeměpis	jiné	ne	80	a
3	26	541	571	1	ZSV	dějepis	ano	50	a
4	4	185	177	0	ZSV	jiné	ano	50	a
5	21	135	180	0	ZSV	jiné	ano	55	a
6	34	147	151	0	ZSV	dějepis	ano	40	a
7	32	0	0	0	fyzika	biologie	ne	70	b
8	5	0	0	1	jiné		ne	90	b
9	11	0	0	0	jiné		ne	40	b
10	24	60	90	1	ZSV	zeměpis	ano	63	b
11	10	0	0	1	ZSV	matematika	ano	69	b
12	23	420	449	0	ZSV	jiné	ano	60	b
13	2	50	50	0	ZSV	zeměpis	ano	60	b
14	25	56	67	0	ZSV	dějepis	ano	65	b
15	40	140	132	0	ZSV	dějepis	ano	70	b
16	12	0	0	0	ZSV	dějepis	ano	80	b
17	19	506	475	0	ZSV	dějepis	ano	51	b
18	8	0	0	0	ZSV	dějepis	ano	45	b
19	35	110	120	0	ZSV	chemie	ano	50	b
20	1	100	97	0	ZSV	jiné	ano	60	b
21	15	60	67	0	ZSV	jiné	ano	55	b
22	30	1691	1670	1	dějepis		ne	50	c
23	20	140	135	1	matematika	jiné	ne	70	c
24	14	0	0	0	zeměpis	jiné	ne	40	c
25	17	117	138	1	ZSV	matematika	ano	65	c
26	16	250	282	0	ZSV	dějepis	ano	50	c

6.Na jaký výsledek v kurzu si věříte?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	a	6	15,0	15,0	15,0
	b	15	37,5	37,5	52,5
	c	12	30,0	30,0	82,5
	d	5	12,5	12,5	95,0
	e	2	5,0	5,0	100,0
	Total	40	100,0	100,0	

Jaký je medián proměnné Q7_stres1?

Kardinální proměnné

- Intervalové a poměrové (SPSS nerozlišuje – obě jsou *scale*)
- Numerické kódy (zpravidla) odpovídají reálným pozorovaným hodnotám
- Více možností jednorozměrné analýzy oproti nominálním a ordinálním proměnným

Co můžeme dělat s kardinálními proměnnými

- Modus a četnosti
  - Udělat můžeme
  - ale při velkých vzorcích nebo velké variabilitě proměnné nejsou užitečné
  - ALE … viz příští hodina o grafech
- Minimum a maximum
- Medián
- Průměr
- Rozptyl/směrodatná odchylka
- kvantily

Míry centrální tendence

- „typická“ hodnota
- Nejlepší reprezentant proměnné

- Použití závisí na typu proměnné:
  - Nominální – modus
  - Ordinální – modus, medián
  - Kardinální – modus, medián, průměr

Průměr

- Aritmetický průměr = součet hodnot / počet případů
- Stejná vzdálenost k případům s nižšími hodnotami jako k případům s vyššími hodnotami
- Citlivý na extrémní hodnoty
- Průměrná mzda vs. mediánová mzda

Měsíční příjmy hostů restaurace v tis. Kč

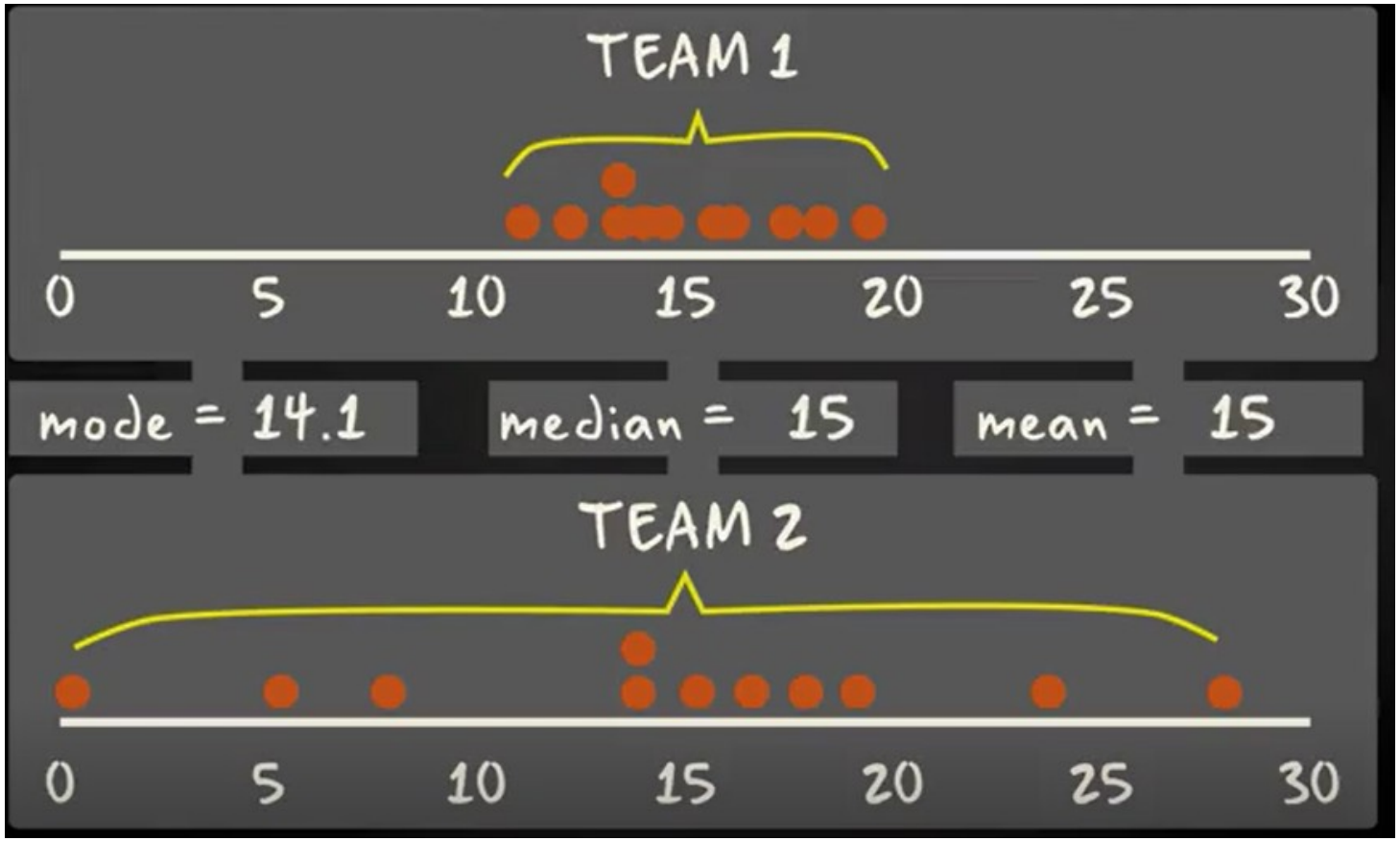
- **Příklad 1:**
  - 11 hostů: 20, 30, 35, 40, 45, 50, 55, 60, 70, 75, 80
  - Medián = 50k
  - Průměr = 50,9k

- **Příklad 2:**
  - 13 hostů: 20, 30, 35, 40, 45, 50, 55, 60, 70, 75, 80, 400, 450
  - Medián = 55k
  - Průměr = 108,5k

- **Příklad 3:**
  - Do restaurace vstoupí Elon Musk a Bill Gates
  - Medián = ?
  - Průměr = ?

Míry centrální tendence

- Užitečné ukazatele, někdy však nemusí stačit
- Např. dva soubory dat mají stejné průměry, ale ve skutečnosti se dost odlišují
- Důležité je znát i míru rozptýlení dat (dispersion)



Rozptyl (variance)

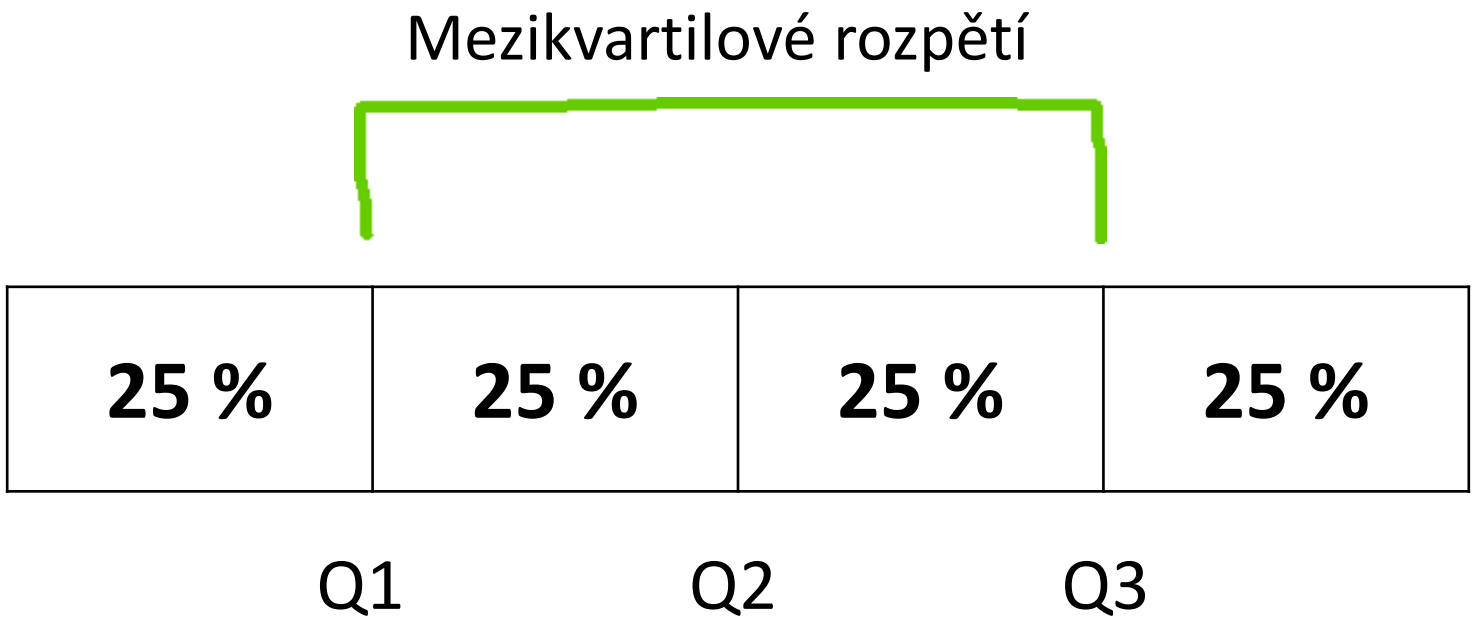
- Suma umocněných odchylek od průměru
  - 9 hostů: 20, 35, 40, 45, 50, 55, 60, 65, 80
  - Průměr = 50k

  - (20-50)+(35-50)+(40-50)+(45-50)+(50-50)+(55-50)+(60-50)+(65-50)+(80-50)
  - -30 + - 15 + -10 + -5 + 0 + 5 + 10 + 15 + 30
  - 900 + 225 + 100 + 25 + 0 + 25 + 100 + 225 + 900
  - 2500/(n-1) = 312,5

  - **Směrodatná odchylka** = odmocnina z rozptylu
    - Průměrná odchylka od průměru
    - Bude **velice důležitá** v dalších hodinách !!!

Mezikvartilové rozpětí

- Interquartile range (IQR)
- Umožňuje snížit citlivost na odlehlé případy
- Kvartily – hodnoty, které rozdělují soubor dat na 4 stejně velké skupiny = „poloviční mediány“
- První kvartil (Q1), druhý kvartil (Q2), třetí kvartil (Q3)



- **IQR = Q3 – Q1**

- Co je Q2?

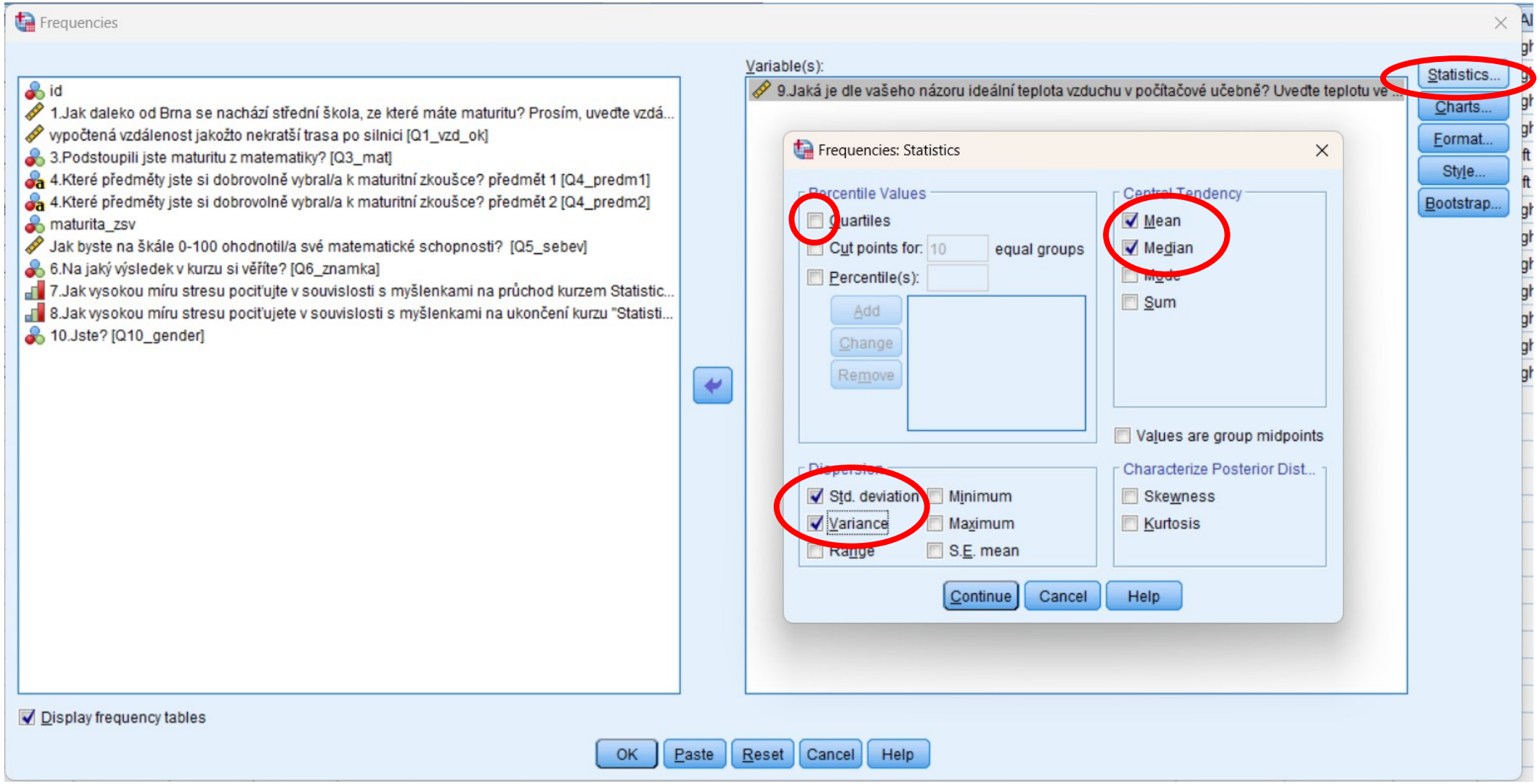
Mezikvartilové rozpětí - postup

Najdeme Q1 a Q3

IQR = Q3 – Q1 = 31,2 – 4,3 = **26,9**

V spss

- Pro proměnnou Q9_teplo
- Analyze → Descriptive Statistics → Frequencies → Statistics



A co pro proměnnou Q1_vzd_ok