Deskriptivní statistika (kategorizované proměnné)

Nejprve malé opakování:

- **Deskriptivní statistika** se zabývá popisem dat, jejich sumarizaci a prezentací.
- **Kategorizované proměnné** jsou všechny proměnné, jejichž hodnoty se nacházejí v určitých kategoriích. Jedná se tedy o nominální, ordinální i kardinální proměnné (pouze ale kardinální poměrové).

Různé druhy proměnných umožňují různé druhy popisu.

Popis nominálních proměnných

U nominálních proměnných zjišťujeme:

- **rozložení četností** variant znaku (pomocí tabulek četností),
- nejčastěji zastoupenou kategorii – **modus** (modálních kategorií někdy může být více než 1),
- **variační poměr**, který se vypočítá tak, že od jedné odečteme podíl četnosti modální kategorie a velikosti souboru.

Popis ordinálních proměnných

U ordinálních proměnných zjišťujeme:

- rozložení četností variant znaku (pomocí tabulek četností),
- nejčastěji zastoupenou kategorii – **modus** (modálních kategorií někdy může být více než 1),
- **medián** (mediánovou kategorii),
- variační poměr,
- další vlastnosti, kterými se ale nebudeme dopodrobna zabývat.

Popis a kontrola dat

Prvním úkolem výzkumníka je popis výběrového souboru. Charakteristikou vzorku by měla začít každá analýza i analytická kapitola v bakalářské či diplomové práci. Zajímá nás například:

- Kolik je ve výběrovém souboru jednotek?

- Kolik je v souboru mužů a žen?
- Kolik je v souboru lidí se ZŠ/SŠ/VŠ vzděláním?
- Jak je v souboru distribuován věk?

Toto rozložení může být vyjádřeno v **absolutních**, **relativních**, či **kumulativních relativních četnostech**.

- **Absolutní četnost** udává absolutní číslo – hodnotu četnosti varianty proměnné v souboru.
  Například: V souboru je 1456 mužů a 1201 žen.
- **Relativní četnost** udává **podíl** četnosti varianty proměnné v souboru.
  Například: V souboru je 24 % osob se základním vzděláním.
- **Kumulativní relativní četnost** udává kumulativní podíly variant proměnné v souboru (nejsou použitelné pro nominální proměnné).
  Například: V souboru je 36 % respondentů, kteří mají alespoň maturitu (tedy nejen úspěšní středoškoláci s maturitou, ale také vysokoškoláci se všemi variantami diplomů).

Popis a kontrola kategorizovaných dat

Tabulky četností

Pro zobrazení základních hodnot popisu rozložení hodnot kategorizovaných proměnných (tedy proměnných nominálních a ordinálních s menším počtem variant odpovědí) se používá tzv. **tabulka četností**. Ta obsahuje jak absolutní, tak relativní četnosti hodnot proměnných. Takto vypadá správná a kompletní tabulka četností:

Jaké je Vaše vzdělání?				
		Četnost odpovědí	**Relativní četnost**	**Validní relativní četnost**
Validní hodnoty	Základní	46	7,5 %	7,6 %
	Základní vyučen /střední bez maturity	62	10,1 %	10,2 %
	Střední s maturitou	307	50,1 %	50,5 %
	Pomaturitní nástavba, VOŠ	40	6,5 %	6,6 %
	Vysokoškolské	153	25,0 %	25,2 %
	Celkem validní hodnoty	608	99,2 %	100,0 %
Chybějící hodnoty (neví, neodpověděl/a)	Chybějící hodnoty	5	0,8 %	
Celkem		**613**	**100,0 %**	

V praxi se často používá jen zkrácená verze tabulky obsahující pouze validní četnosti:

Jaké je Vaše vzdělání?	**Četnost odpovědí**	**Validní relativní četnost**
Základní	46	7,6 %
Základní vyučen /střední bez maturity	62	10,2 %
Střední s maturitou	307	50,5 %
Pomaturitní nástavba, VOŠ	40	6,6 %
Vysokoškolské	153	25,2 %

Celkem	**608**	**100,0 %**

Před počítáním četností je ale potřeba zkontrolovat data. Kontrolujeme, zda se nachází v platném intervalu (například proměnná pohlaví nabývá v našem souboru pouze hodnot 1 a 2, všechny ostatní varianty by měly být omyly).

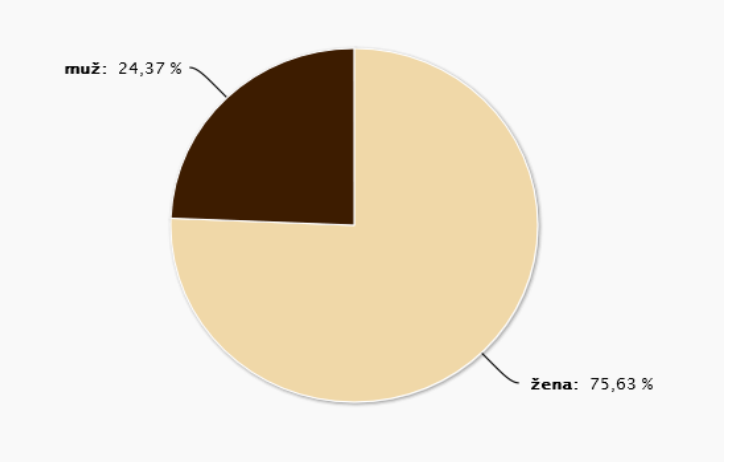
Grafy četností

Pro znázornění rozložení četností se využívají i grafy znázorňující četnosti hodnot proměnných. Nejznámějšími variantami jsou koláčový a sloupcový graf.

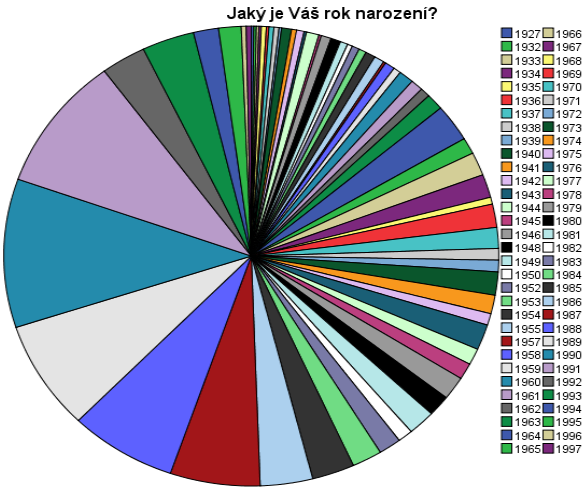
Koláčový graf je vhodný:

- pro třídění prvního stupně (jedna datová řada),
- pro porovnání četností u nominálních proměnných, které nemají příliš mnoho hodnot (méně než 7),
- pokud hodnoty, které chcete vykreslit, nejsou nulové,
- pokud hodnoty představují část celku.

Příklad proměnné, kde je vhodné využít koláčový graf:



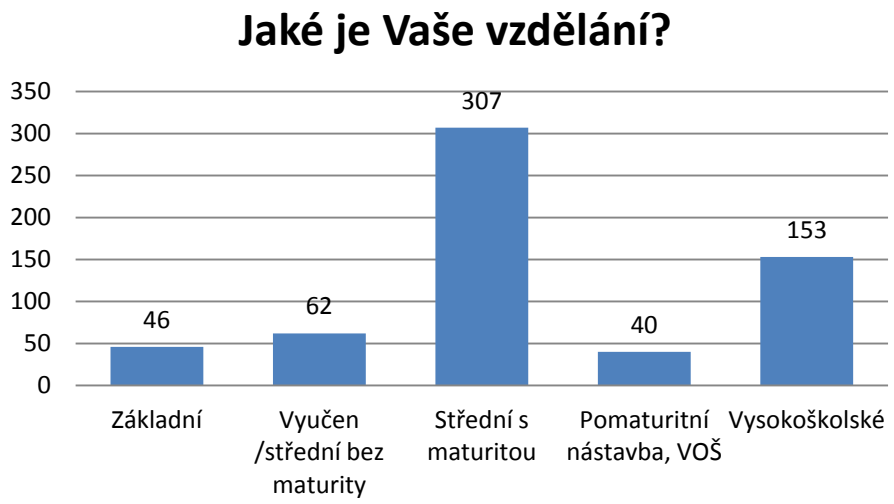
Příklad proměnné, kde NENÍ vhodné využít koláčový graf:



Sloupcový graf je vhodný pro:

- porovnání položek,
- ordinální proměnné a kardinální proměnné s menším počtem kategorií,
- znázornění změn za časové období (třídění druhého stupně).

Příklad sloupcového grafu:

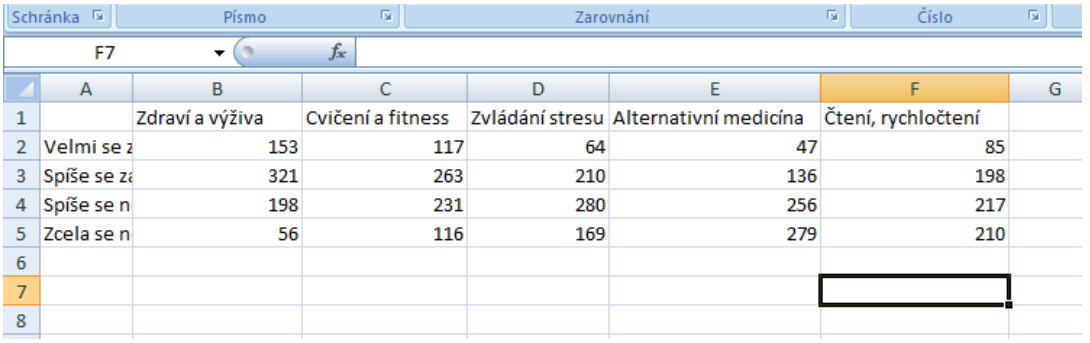


Grafy se v Excelu vkládají pomocí funkce „**Grafy**“ na listu „**Vložení**“.

Porovnání rozložení četností

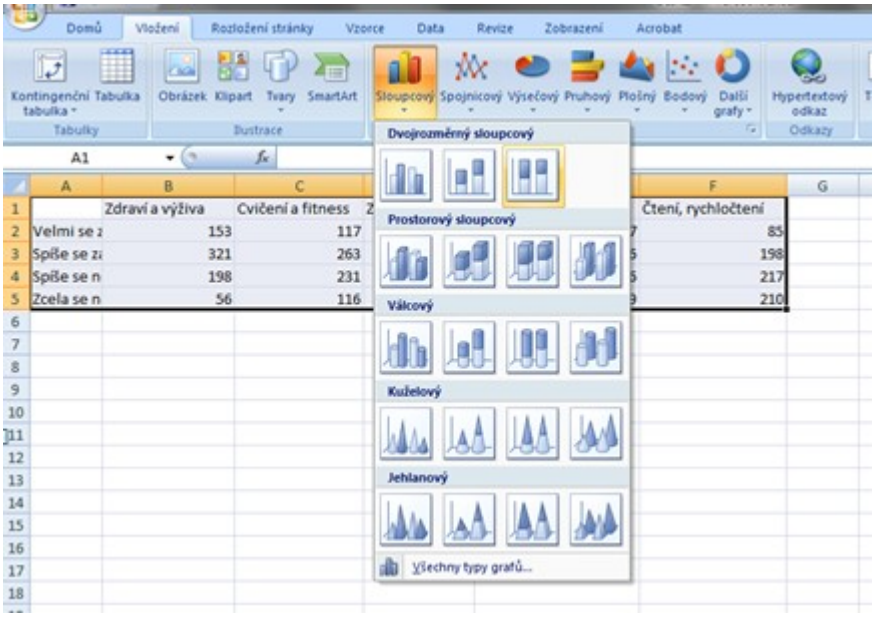
Pro zobrazení porovnání rozložení četností u baterií otázek se používají **skládané sloupcové grafy**.

Skládaný sloupcový graf můžete vytvořit tak, že si připravíte tabulku s absolutními validními četnostmi u jednotlivých kategorií:

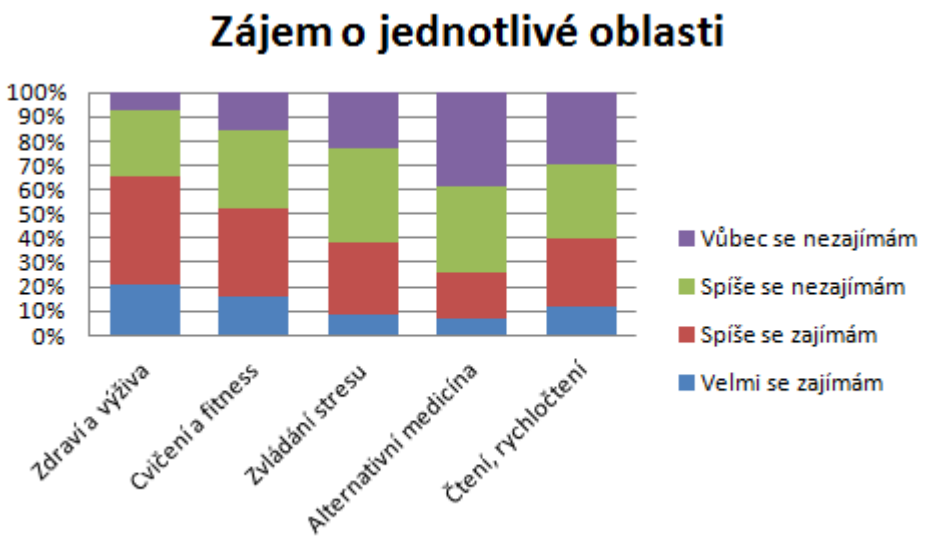


	A	B	C	D	E	F	G
1		Zdraví a výživa	Cvičení a fitness	Zvládání stresu	Alternativní medicína	Čtení, rychločtení	
2	Velmi se z	153	117	64	47	85	
3	Spíše se za	321	263	210	136	198	
4	Spíše se n	198	231	280	256	217	
5	Zcela se n	56	116	169	279	210	
6							
7							
8							

Tabulku si označíte a zvolíte možnost „Vložení“ – „Grafy“ – „Sloupcový“.



Výsledkem je skládaný sloupcový graf, který přehledně ukazuje rozdíly v rozložení jednotlivých proměnných.



Modus a medián

Pro připomenutí z minulého semestru si uveďme, v čem se liší MODUS a MEDIÁN (obě udávají tzv. míry centrální tendence a často se pletou):

> ***MODUS*** *je hodnota, která se v datech vyskytuje nejčastěji.*
> ***MODÁLNÍ KATEGORIE*** *je tedy nejpočetněji zastoupená kategorie.*
>
> ***MEDIÁN*** *dělí řadu výsledků seřazených podle velikosti na dvě stejně početné poloviny.*
> ***MEDIÁNOVÁ KATEGORIE*** *je ta, ve které je dosaženo 50% všech údajů, postupujeme-li od první kategorie výše.*

Jestliže je počet položek ve výzkumném souboru lichý, pak platí:

$$\textbf{Medián} = x_{(n+1)/2)}$$

Jestliže je počet položek ve výzkumném souboru sudý, pak platí:

$$\textbf{Medián} = 0{,}5(x_{n/2}+x_{n/2+1})$$

Představte si otázku na počet dětí. Odpovědi respondentů jsou {0,1,1,2,2,3,5}.

- *V souboru jsou dvě modální kategorie (tedy kategorie s nejvyšším počtem výskytů) – jsou to hodnoty 1 a 2.*
- *Mediánová kategorie je 2. Medián je na rozdíl od aritmetického průměru málo citlivý k odlehlým (extrémním) hodnotám. Pokud by byly odpovědi respondentů {0,1,1,2,2,3,5,10}, medián stale zůstává roven 2.*

Modus a medián v Excelu

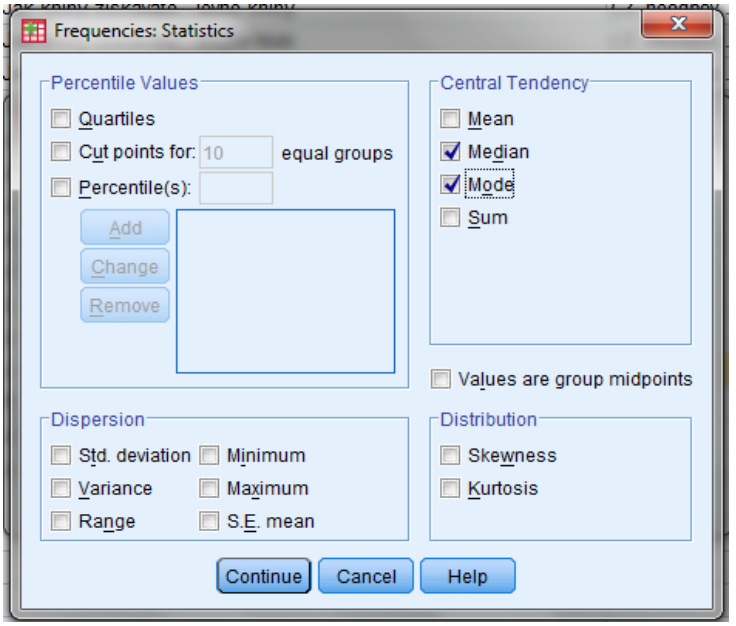
V Excelu existují na výpočet mediánu a modu jednoduché příkazy MEDIAN a MODE. Syntaxe zápisu je snadná:

- =MEDIAN(datová oblast) – např. =MEDIAN(A1:A730)
- =MODE(datová oblast) – např. =MODE(A1:A730)

(Příkazy vypočítají medián a modus ze sloupce A, řádků 1-730.)

Modus a medián v SPSS

V SPSS vyberete v nabídce položky Analyze > Descriptive Statistics > Frequencies (zde zvolíte proměnnou) > Statistics > Median, Mode.



Tipy pro vytváření grafů

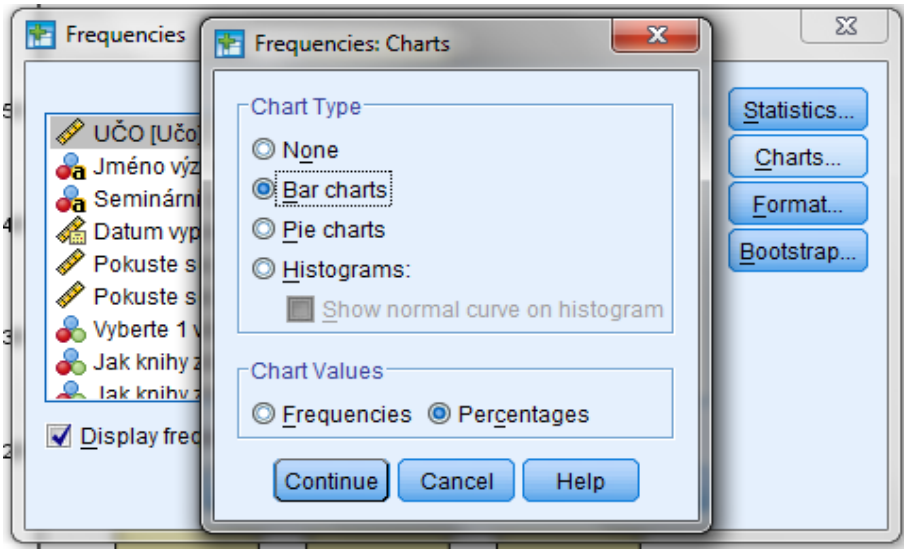
Levine a Stephan (2010) shrnují několik tipů pro prezentaci dat prostřednictvím grafů v akademickém prostředí:

- vždy si vyberte ten nejjednodušší graf,
- vždy používejte popisek grafu,
- popište obě osy,
- vyvarujte se ilustrací a zbytečného používání grafiky na pozadí nebo okrajích grafu,
- vyvarujte se používání módních piktogramů, které by mohly ztížit čitelnost dat,
- vertikální osa by měla začínat nulou (pokud nezačíná negativními hodnotami).

V neakademickém prostředí (např. pro účely marketingu) je využití grafiky vhodné, v prostředí akademickém je na prvním místě čitelnost dat. 3D efekty a vkládání obrázků mohou znemožnit čtení hodnot dat. Další tipy pro vytváření grafů najdete třeba zde.

Tabulky četností a grafy v SPSS

Tabulky četností v SPSS získáme příkazem Analyze → Descriptive Statistics → Frequencies . Grafy vytvoříme cestou Analyze → Descriptive Statistics → Frequencies → **Charts**.



Spojité proměnné

Spojité (nekategorizované) proměnné jsou ty proměnné, které mohou nabývat všech hodnot z daného intervalu. Může jednat o plat, věk, počet obyvatel města, délku pracovní zkušenosti v měsících…

Aritmetický průměr

Aritmetický průměr je třetí mírou centrální tendence. U kardinálních dat lze jako míry centrální tendence využívat všechny tři:

- modus,
- medián,
- aritmetický průměr.

Aritmetický průměr je ukazatelem „průměrné“ hodnoty, nemusí být ale vždy ukazatelem nejvhodnějším – vhodné je jej kombinovat s mediánem. Aritmetický průměr je totiž velmi citlivý na extrémní hodnoty. I jedna extrémní hodnota může výrazně posunout aritmetický průměr.

Příklad: V roce 2010 byl podle serveru Platy.cz průměrný měsíční plat 23 300 Kč. Medián byl však na hodnotě 21 000 Kč. Znamená to, že průměr vychýlil menší počet jedinců s výrazně vyšším platem.

Průměrný měsíční plat (v Kč)	Medián (Kč)	Rozdíl (v %)
23 300	21 000	11%

Zdroj: Platy.cz

Pro připomenutí:

Modus se používá, pokud:

- rozdělení má více vrcholů,
- chceme zjistit nejčastější hodnoty.

Medián používáme, pokud:

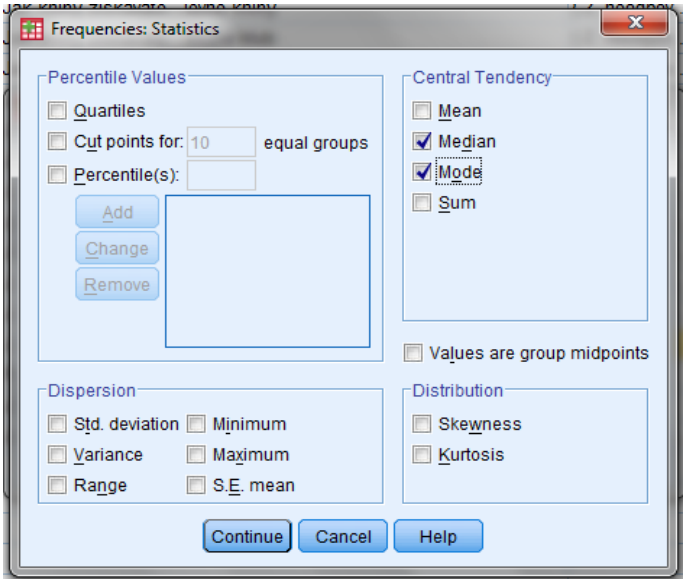
- jsou data ordinální nebo kardinální,
- chceme znát střed rozložení dat,
- (v kombinaci s průměrem) pokud soubor obsahuje extrémní hodnoty,
- jestliže je rozložení dat zešikmené.

Aritmetický průměr je vhodné používat, pokud

- jsou data kardinální,
- rozložení je symetrické,
- chceme použít statistické testy. (Hendl 2009)

Aritmetický průměr v SPSS

Pro zjištění hodnot měr centrální tendence v SPSS zadáte Analyze → Descriptive Statistics → Frequencies → **Statistics → Mean, Median, Mode**



Minimum, maximum a rozpětí

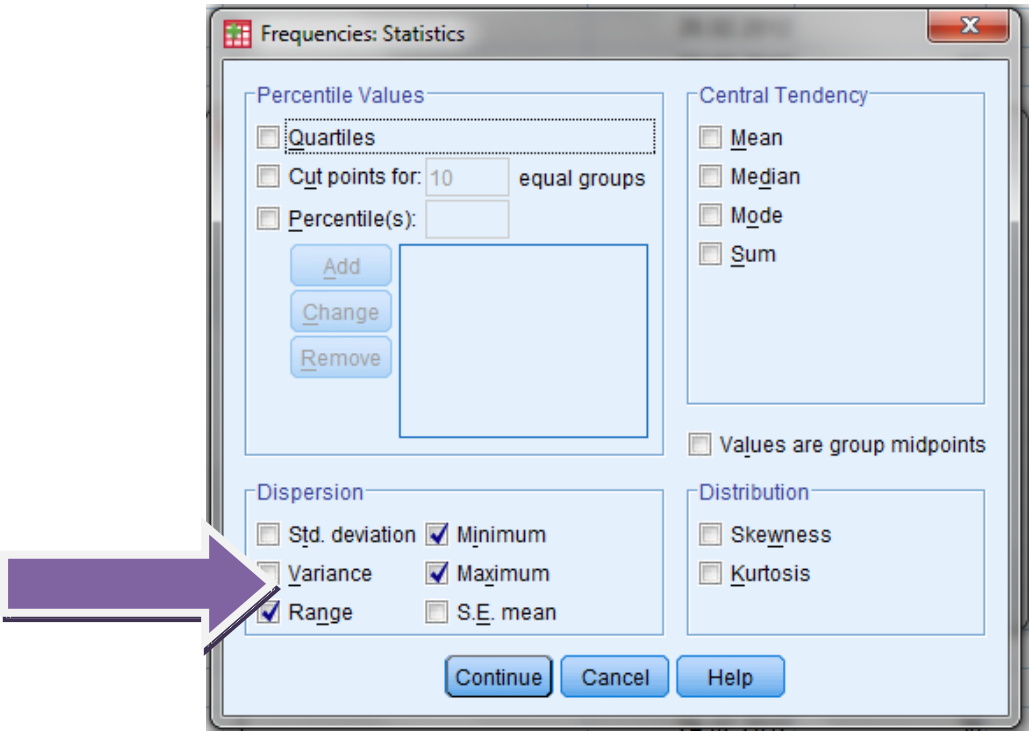
První charakteristiky nekategorizovaných dat, na které se díváme už při fázi čištění dat, jsou **minimální** a **maximální hodnoty**. Z nich také snadno spočítáme **rozpětí**.

Rozpětí je nejjednodušší míra variability a snadno se vypočítá jako rozdíl mezi nejvyšší a nejnižší hodnotou.

> *Např. Je-li minimální hodnota 18 a maximální 1024, rozpětí hodnot proměnné v souboru je 106.*

Minimum, maximum a rozpětí v SPSS

Vypočítání rozpětí můžete v SPSS zadat tímto řetězcem: **Analyze – Frequencies – Statistics:**



Rozptyl a směrodatná odchylka

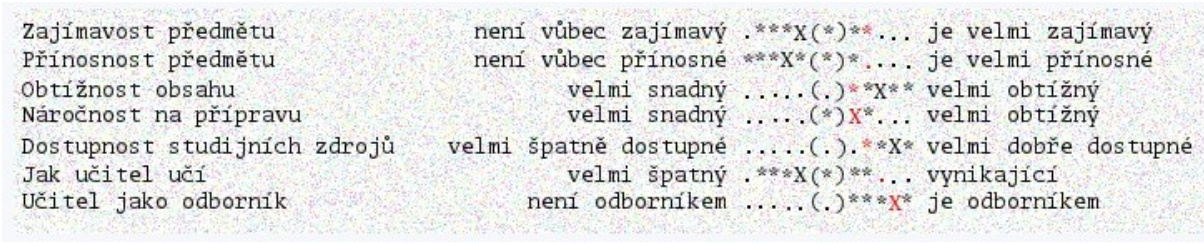
Rozptyl je definován jako střední hodnota kvadrátů odchylek od střední hodnoty (průměru). Vyjadřuje variabilitu rozdělení souboru náhodných hodnot kolem její střední hodnoty. Při průměrování odchylek dělíme číslem n-1.

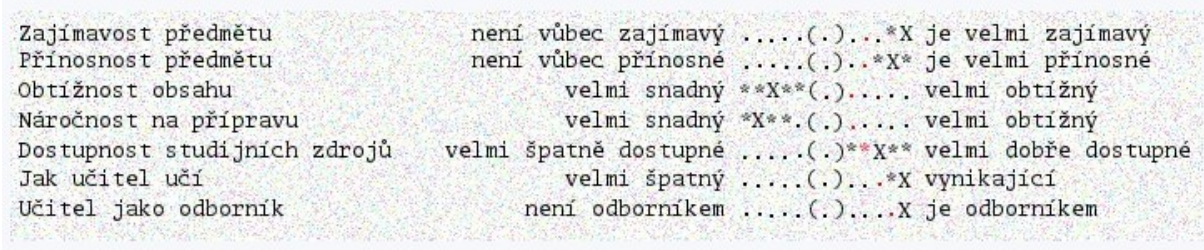
S rozptylem úzce souvisí **směrodatná odchylka**. Ta se vypočítá jako odmocnina z rozptylu. Vrací tedy míru rozptýlenosti do měřítka původních dat. V podstatě nám říká, uvnitř jakého intervalu okolo průměru leží zvolené procento případů – tedy čím je směrodatná odchylka menší, tím lépe pro aritmetický průměr.

Hendl (2009) srozumitelně vysvětluje, jak dochází k výpočtu směrodatné odchylky:

1. Nejprve si vypočítáme všechny odchylky od průměru (např. při hodu kostkou vždy spočítáme odchylku konkrétní hozené hodnoty od celkového průměru).
2. Umocnění na druhou převede záporné odchylky na kladná čísla. Zároveň zvýrazní váhu extrémnějších odchylek.
3. Sečteme kvadratických odchylek.
4. Dělením číslem n-1 získáme průměrnou kvadratickou odchylku.
5. Odmocnina (v případě směrodatné odchylky) převede výsledek do původního měřítka dat.

Pro názornost si pojďme ukázat příklad, který dobře znáte – hodnocení vyučujících na KISKu a směrodatnou odchylku tohoto hodnocení.





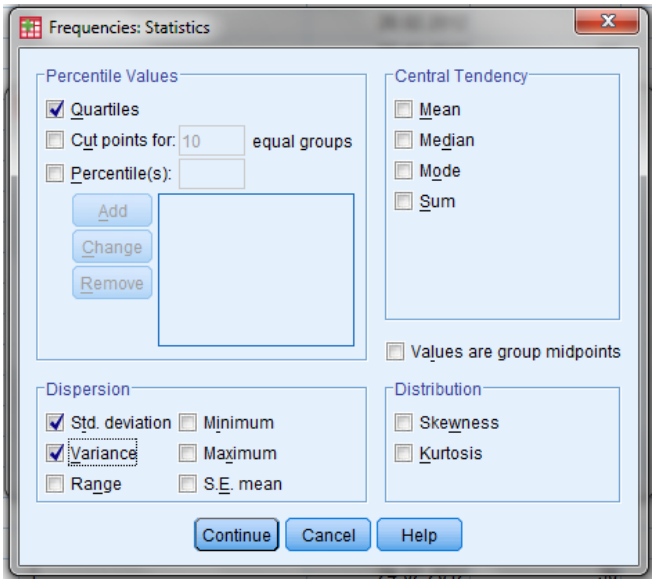
Průměrné hodnocení proměnné „Učitel jako odborník" je u obou vyučujících podobné – jeden vyučující má průměrné hodnocení 9, druhý má průměrné hodnocení 10. Směrodatná odchylka (zvýrazněná hvězdičkami) nám ale poskytne rychlou další informaci – říká nám, jak moc se hodnocení všech respondentů pohybovalo kolem průměru. Vidíme, že zatímco v druhém případě se hodnocení výjimečně shodovalo a studující se shodli na tom, že učitel je skutečný odborník, v prvním případě nebyla shoda zdaleka tak veliká.

Rozptyl a směrodatná odchylka v Excelu

- rozptyl – příkaz **VAR**
- směrodatná odchylka – příkaz **SMODCH.VÝBĚR**

Rozptyl a směrodatná odchylka v SPSS

Vypočítání rozptylu a směrodatné odchylky můžete v SPSS zadat tímto řetězcem:
Analyze – Frequencies – Statistics:



Percentily

Percentil x je hodnota, pro kterou platí, že x procent případů má hodnotu menší nebo rovnu percentilu x.

Nejčastěji se využívají:

- **MEDIÁN** (x50)
- **KVARTILY** (x25, x50, x75)
- **DECILY** (x10, x20, x30, x40, x50, x60, x70, x80, x90)

Například vás může zajímat, jak jsou rozloženy příjmy obyvatel v horním a spodním percentilu. Tato informace spolu s mediánem ukazuje, jak moc jsou rozevřené pomyslné nůžky mezi „horní“ a „spodní“ vrstvou společnosti.

Jak vysoký je medián proti průměrné mzdě? (ve vybraných zemích OECD)

Země	spodních 10 %	medián	horních 10 %
Švédsko	56 %	89,8 %	150,9 %
Finsko	62,3 %	89,5 %	147,9 %
Kanada	44,6 %	89,1 %	166,9 %
Dánsko	60,9 %	89 %	150,4 %
Norsko	63,2 %	88,9 %	149 %
Japonsko	52,4 %	87,6 %	162,7 %
Nový Zéland	51,2 %	87,2 %	160,6 %
Německo	43,4 %	87 %	165,7 %
Česko	**49,3 %**	**85,2 %**	**153,1 %**
Itálie	56,1 %	85,1 %	156,6 %
Švýcarsko	56,6 %	84,9 %	153,4 %
Belgie	60,4 %	84,5 %	153,4 %
Nizozemí	51,7 %	84 %	158,8 %

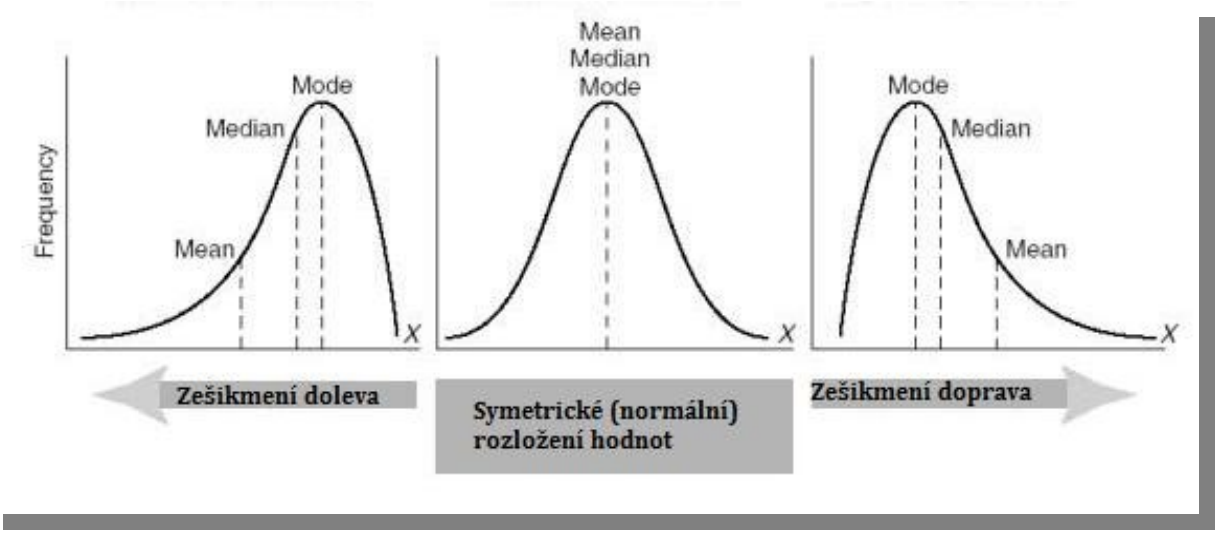
Zdroj: http://finexpert.e15.cz/jak-se-lisi-prumerna-mzda-a-median

Percentil v SPSS

Vypočítání rozptylu a směrodatné odchylky můžete v SPSS opět zadat tímto řetězcem: **Analyze – Frequencies – Statistics (políčko Percentile Values)**.

Šikmost a špičatost

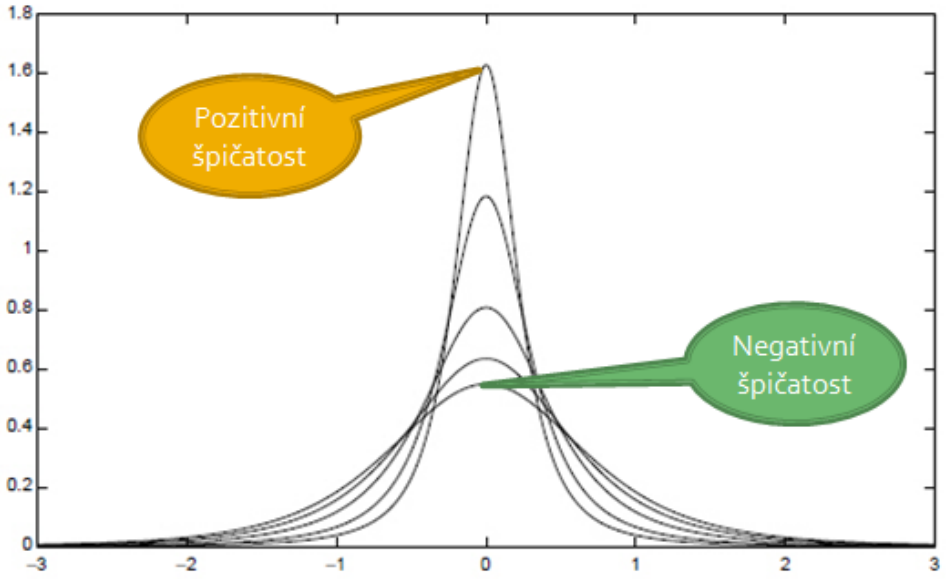
Spojitá data nabývají málokdy tzv. normálního rozložení. Při popisu dat si všímáme zešikmení a špičatosti dat.



Ad **šikmost**:

- **Symetrické (normální) rozložení** - aritmetický průměr, medián a modus mají stejné nebo velmi podobné hodnoty. (0)
- Pokud je aritmetický průměr větší než medián, který je zase větší než modus, znamená to, že je více případů menších než průměr a naše **rozložení je šikmé doprava**. (+)
- Třetí možností je, že je více případů větších než aritmetický průměr. Ten je pak menší než medián a ten je menší než modus. Naše **rozložení je šikmé doleva**. (-)

Špičatost zase udává, jak moc jsou data nakumulována v oblasti středních hodnot.



Šikmost a špičatost v SPSS

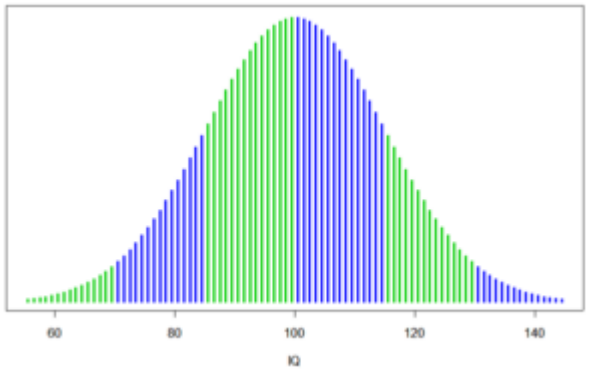
Analyze – Frequencies – Statistics (políčko Distribution).

Zobrazování kardinálních dat

Pro zobrazování kardinálních dat se používá několik možných grafů

Histogram

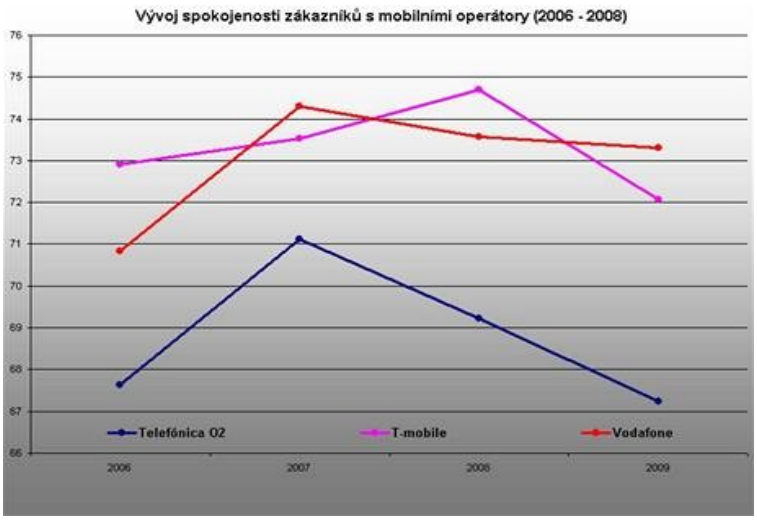
Histogram je podobný sloupcovému grafu, mezi jednotlivými sloupci ale nejsou mezery. Pracujete-li v Excelu, můžete využít klasický sloupcový graf.



Příklad histogramu – distribuce IQ v populaci (zdroj: IQscope.com)

Spojnicové grafy

Chcete-li ukázat, jak se hodnoty proměnné měnily v čase, je vhodné použít spojnicový graf.



Příklad využité spojnicového grafu – spokojenost s mobilními operátory 2006-2008

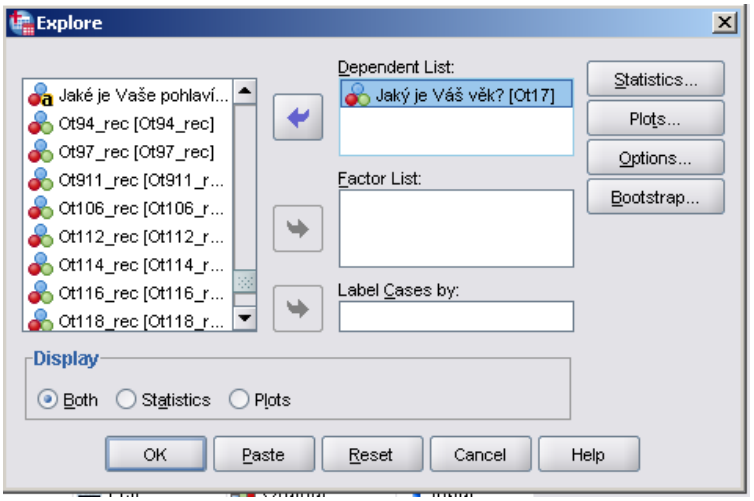
Bodové grafy

Bodové grafy zachycují jednotlivé hodnoty proměnných a využívají se v třídění druhého stupně jako zachycení toho, jak jedna proměnná ovlivňuje druhou (o tomto grafu více v dalších modulech).

Procedura EXPLORE

Ke zpracování kardinálních proměnných se hodí procedura EXPLORE

Nevytváří tabulky hodnot proměnné (jako frequencies) ale zobrazuje souhrnné statistiky a grafy - Zadání: **ANALYZE -> DESCRIPTIVE STATISTICS -> EXPLORE**



V „Plots“ je možné nastavovat další grafy. Standardně je výstupem tzv. Box-plot a Stem and Leaf (ten je ale nepříliš přehledný). Lze také zvolit histogram (není součástí default nastavení)

Výstupy z EXPLORE:

Tabulka s hlavními statistikami:

Pro průměr máme vypočtenou také standardní chybu, která vyjadřuje spolehlivost dat. Za předpokladu prostého náhodného výběru bychom mohli říci, že průměr v základní populaci (z níž je náš soubor vzorkem) by ležel s 95% pravděpodobností v intervalu +- 2 standardní chyby průměru. Tedy zhruba mezi 22,99 a 23,29 roku. Všimněte si, že interval spolehlivosti je velmi úzký – je to tím, že pracujeme s velmi rozsáhlým souborem – při velikosti téměř 3000 respondentů je výběrová chyba poměrně malá.

Dále máme tzv. robustní průměr – bez 5 % odlehlých případů. (například onen 100letý šprýmař zde vypadl a průměr se snížil)

Vedle tzv. měr centrální tendence (průměr, medián příp. modus) stojí u kardinálních proměnných vždy za povšimnutí míry variability. Rozptyl a směrodatná odchylka poukazují na to, jak moc jsou data rozházená kolem průměru. Malá hodnota = všichni v podobném věku, velká hodnota = velice rozmanité stáří studentů)

Descriptives

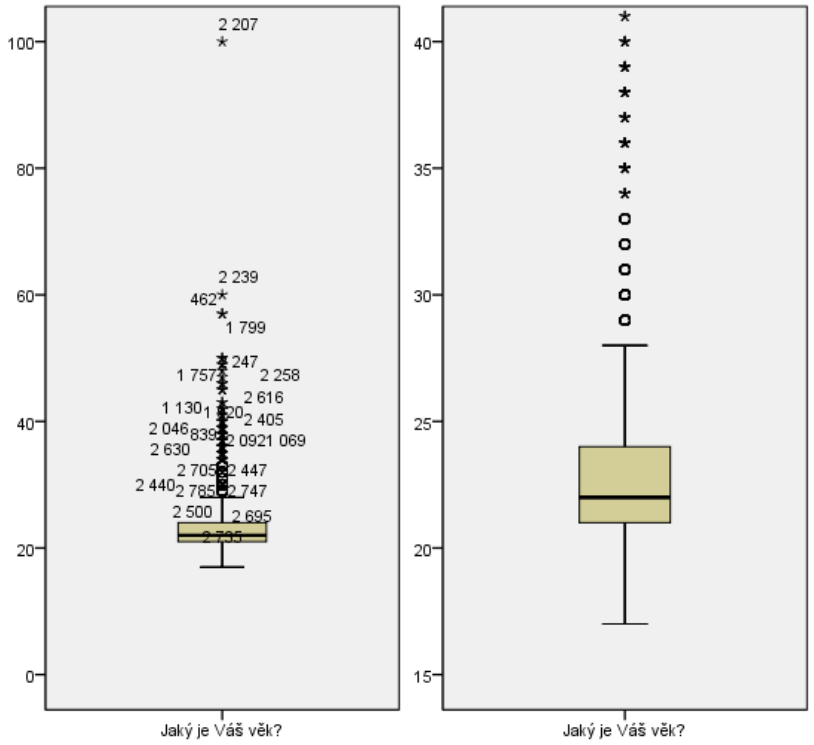
			Statistic	Std. Error
Jaký je Váš věk?	Mean		23,14	,076
	95% Confidence Interval for Mean	Lower Bound	22,99	
		Upper Bound	23,29	
	5% Trimmed Mean		22,68	
	Median		22,00	
	Variance		16,100	
	Std. Deviation		4,012	
	Minimum		17	
	Maximum		100	
	Range		83	
	Interquartile Range		3	
	Skewness		5,190	,047
	Kurtosis		61,778	,093

Explore umí: Krabicový graf - je užitečný:

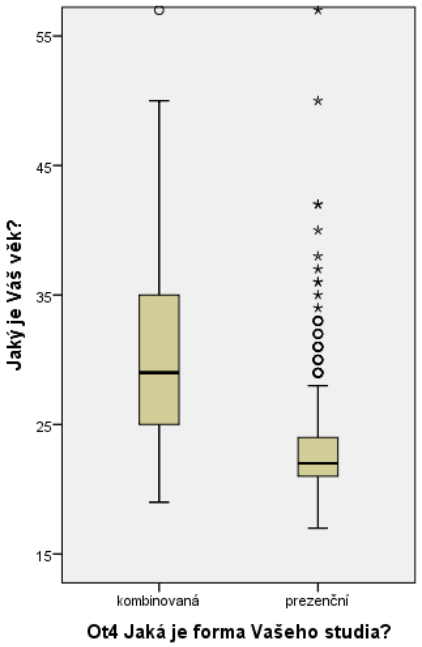
- když chceme rychle identifikovat odlehlé případy (čísla případů jsou přímov grafu – pak by stačilo najít daný řádek v matici a zkontrolovat, zda respondent nevyplnil podivně i jiné otázky, nebo nejde o překlep)
- když chceme získat rychlý přehled o rozložení hodnot (ne jen průměr či medián, ale i to, zda je rozložení symetrické, špičaté, šikmé…)
- když chceme porovnat více rozložení navzájem (třeba věk podle typu studia) Tady už se jedná o dvourozměrnou analýzu a musíme použít kategorizovanou proměnnou vloženou do „factor list“

Vpravo je tentýž graf s rozumným měřítkem. Tlustá čára uprostřed znázorňuje medián, krabice je definována 25. a 75. percentilem. Uvnitř krabice leží 50 % případů a její výška je dána tzv. interkvartilovým rozpětím.

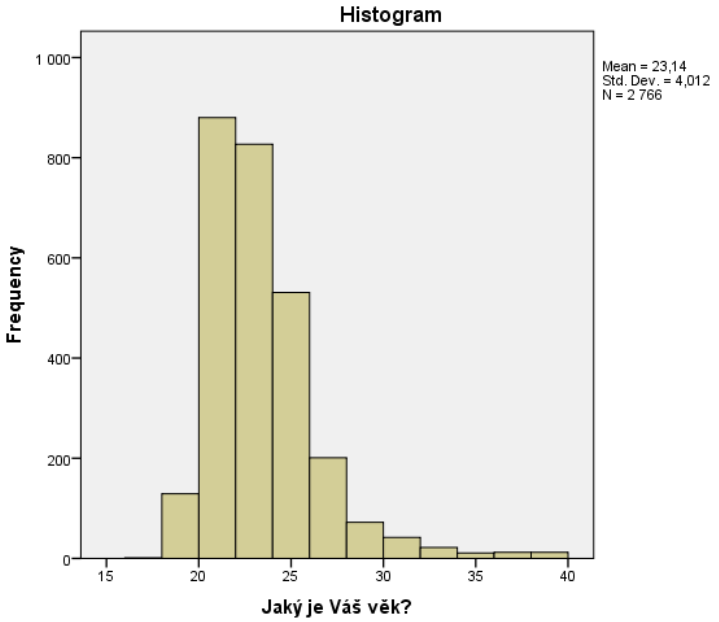
Blíže k tomu viz: http://cs.wikipedia.org/wiki/Boxplot nebo http://www.eistat.cz/popis/boxplot/index.htm



Krabicový graf použitý pro srovnání – vidíme nejen vzájemnou pozici mediánů, ale můžeme srovnat i základní charakteristiky rozložení



Explore umí: Histogram - je ještě přehlednější a zobrazuje celé rozložení proměnné (zde opět měřítko upraveno na 15 – 40 let). Histogram lze získat také v proceduře Frequencies. (od sloupcového grafu se liší tím, že má lineární osu x - nezkresluje tedy vzdálenosti).



Dále je procedura explore užitečná ještě pro zjišťování, zda je proměnná normálně rozložena. K tomu jsou určeny speciální tzv. kvantilové grafy (Q_Q Plot) a testy (kolmogorov-smirnov). Zadat lze v submenu „Plots“ -> „normality plots with tests“

Literatura

Hendl, J. *Přehled statistických metod analýzy dat*. Praha : Portál 2009

Levine, D. M., & Stephan, D. (2010). *Even you can learn statistics: A guide for everyone who has ever been afraid of statistics*. Upper Saddle River, N.J: FT Press.