Analýza dat pro Neurovědy

RNDr. Eva Koriťáková, Ph.D.

doc. RNDr. Ladislav Dušek, Dr.

Přínos kurzu

- Orientace v principech analýzy dat, plánování a hodnocení experimentů z oblasti medicíny.
- Schopnost správné aplikace základních metod analýzy medicínských dat v praxi.
- Schopnost správné interpretace dosažených výsledků.
- Schopnost praktické analýzy dat v softwaru SPSS.

Osnova kurzu

1. Jak medicínská data správně popsat a vizualizovat :
   - Typy dat a jejich vizualizace
   - Předzpracování dat a popisná sumarizace
2. Jak medicínská data správně testovat :
   - Modelová rozdělení dat, transformace dat, intervaly spolehlivosti
   - Formulování hypotéz, hladina významnosti, síla testu, p-hodnota
3. Jak a kdy použít parametrické a neparametrické testy I. :
   - Jednovýběrové testy: z-test, jednovýběrový t-test, párový t-test
   - Dvouvýběrový t-test
   - Neparametrické testy: Wilcoxonův test, Mannův-Whitneyův test
   - F-test
4. Jak a kdy použít parametrické a neparametrické testy II. :
   - Analýza rozptylu (ANOVA) a její předpoklady
   - Problém násobného testování hypotéz – Bonferonniho korekce, FDR
   - Kruskalův-Wallisův test

Osnova kurzu

5. Jak analyzovat kategoriální a binární data I. :
   - Analýza kontingenčních tabulek
   - Relativní riziko (relative risk) a poměr šancí (odds ratio)
   - Binomické a Poissonovo rozdělení
6. Jak analyzovat kategoriální a binární data II. :
   - Hodnocení diagnostických testů – senzitivita, specificita, prediktivní hodnoty
   - Hledání diagnostického cut-off pomocí ROC křivek
7. Jak hodnotit vztah spojitých proměnných a základy regresního modelování :
   - Základy korelační analýzy – Pearsonův a Spearmanův korelační koeficient
   - Základy regresní analýzy – lineární regrese, odstranění vlivu kovariát
8. Jak analyzovat přežití pacientů :
   - Analýza přežití
   - Coxova regrese

Požadavky ke kolokviu

- Předmět je ukončen kolokviem sestávajícím se z analýzy praktických příkladů na počítači.

- Je nutné porozumět probíraným tématům a umět aplikovat základní statistické metody při analýze reálného datového souboru.

Doporučená literatura – v češtině

- Havránek, T., 1993. *Statistika pro biologické a lékařské vědy*. Praha: Academia.
- Benedík, J., Dušek, L., 1993, Sbírka příkladů z biostatistiky. Brno: Konvoj.
- Zvárová, J., 2001. *Základy statistiky pro biomedicínské obory*. Praha: Karolinum. (http://ucebnice.euromise.cz/index.php?conn=0&section=biostat1)

Doporučená literatura – v angličtině

- Zar, J.H., 1998. Biostatistical analysis. London: Prentice Hall.
- StatSoft, Electronic Statistics Textbook (http://www.statsoft.com/textbook/elementary-statistics-concepts/button/1/)
- Harrington, M., 2011. The Design of Experiments in Neuroscience, London: SAGE.
- Weaver, A. & Goldberg, S., 2012. Clinical Biostatistics and Epidemiology Made Ridiculously Simple, Miami: MedMaster.
- Rumsey, D.J., 2010. Statistics Essentials For Dummies, Hoboken: Wiley.
- Rumsey, D.J., 2011. Statistics For Dummies, Hoboken: Wiley.
- Rumsey, D.J., 2009. Statistics II For Dummies, Hoboken: Wiley.
- Salkind, N.J., 2010. Statistics for People Who (Think They) Hate Statistics, London: SAGE.
- Gonick, L. & Smith, W., 2000. The Cartoon Guide to Statistics, London: Harper Collins.
- Oweiss, K.G., 2010. Statistical Signal Processing for Neuroscience and Neurotechnology, Burlington: Academic Press.
- Triola, M.M. & Triola, M.F., 2006. Biostatistics for the Biological and Health Sciences, Boston: Pearson.

Doporučená literatura – workbooky v angličtině

- Rumsey, D.J., 2005. Statistics Workbook For Dummies, Hoboken: Wiley.
- Grove, S.K., 2007. Statistics for Health Care Research: A Practical Workbook, Edinburgh: Elsevier Saunders.
- Petrie, A. & Sabin, C., 2013. Medical Statistics at a Glance - Workbook, Chichester: Wiley-Blackwell.
- Barnette, J.J. & Walters, I.C., 2006. Biostatistics Student's Solutions Manual, Boston: Pearson. (k učebnici Triola & Triola, Biostatistics for the Biological and Health Sciences)

Blok 1

Jak medicínská data správně popsat
a vizualizovat.

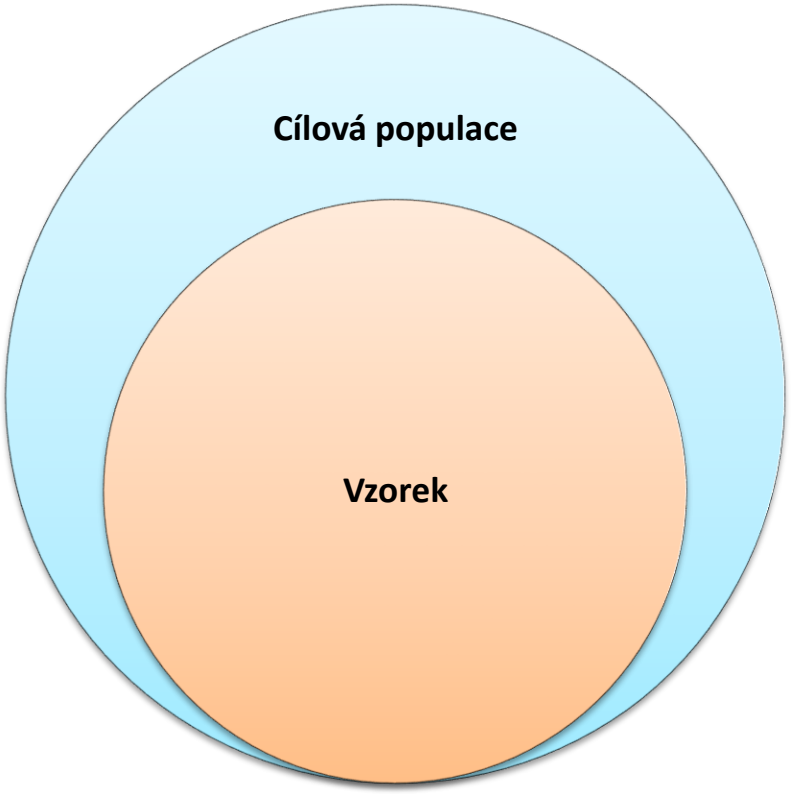
Osnova

1. Typy medicínských dat a jejich vizualizace
2. Předzpracování dat
3. Popisná sumarizace dat

1. Typy medicínských dat a jejich vizualizace

Data

- **Cílová populace** – skupina subjektů, o které chceme zjistit nějakou informaci (např. všichni pacienti s danou diagnózou v ČR).
- Cílová populace = základní soubor
- **Experimentální vzorek** – podskupina (výběr) z cílové populace, kterou „máme k dispozici“ (pozorovaný soubor).
  - Musí odpovídat svými charakteristikami cílové populaci.
  - Chceme totiž zobecnit výsledky na celou cílovou populaci.
- **Data** – číselný nebo slovní záznam informací o pozorovaném souboru lidí, zdravotnických zařízení apod.



Datová tabulka

PROMĚNNÉ

OBJEKTY (SUBJEKTY)

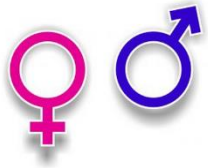
ID	Pohlaví	Věk	Váha	…
1	muž	84	85,5	
2	žena	25	62,0	
3				
4				
…				

Datový soubor – zásady ukládání dat

- Správné a přehledné uložení dat je základem jejich pozdější analýzy.
- Je vhodné rozmyslet si před zahájením sběru dat, jak budou data ukládána.
- Pro počítačové zpracování dat je nezbytné ukládat data v tabulkové podobě:
  - Každý sloupec obsahuje pouze jediný typ dat, identifikovaný hlavičkou sloupce (hlavičky sloupců musejí být unikátní).
  - Každý řádek obsahuje minimální jednotku dat (např. pacient, jedna návštěva pacienta apod.).
  - Je nepřípustné kombinovat v jednom sloupci číselné a textové hodnoty.
  - Komentáře jsou uloženy v samostatných sloupcích.
  - U textových dat je nezbytné kontrolovat překlepy v názvech kategorií.
  - Specifickým typem dat jsou datumy, u nichž je nezbytné kontrolovat, zda jsou uloženy v korektním formátu.

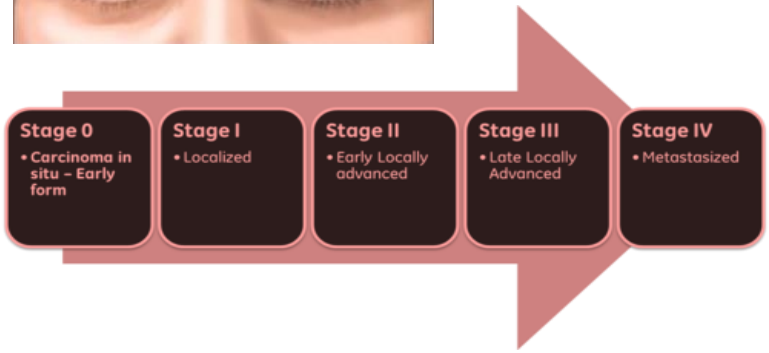
Typy dat

- **Kvalitativní (kategoriální) data:**
  - Binární data



  - Nominální data

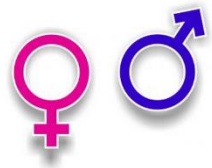
  - Ordinální data



- **Kvantitativní data:**
  - Intervalová data

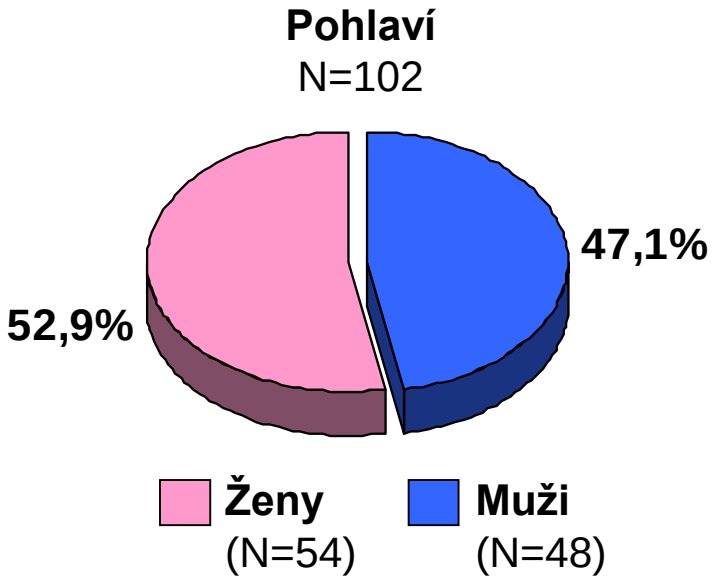
  - Poměrová data

Binární data (kvalitativní)



- Pouze dvě kategorie
- Příklady: pohlaví (muž x žena), onemocnění (ano x ne), kouření (ano x ne)
- Často číselné kódování pomocí 0 (ne) a 1 (ano)
- Rovná se?

Koláčový graf

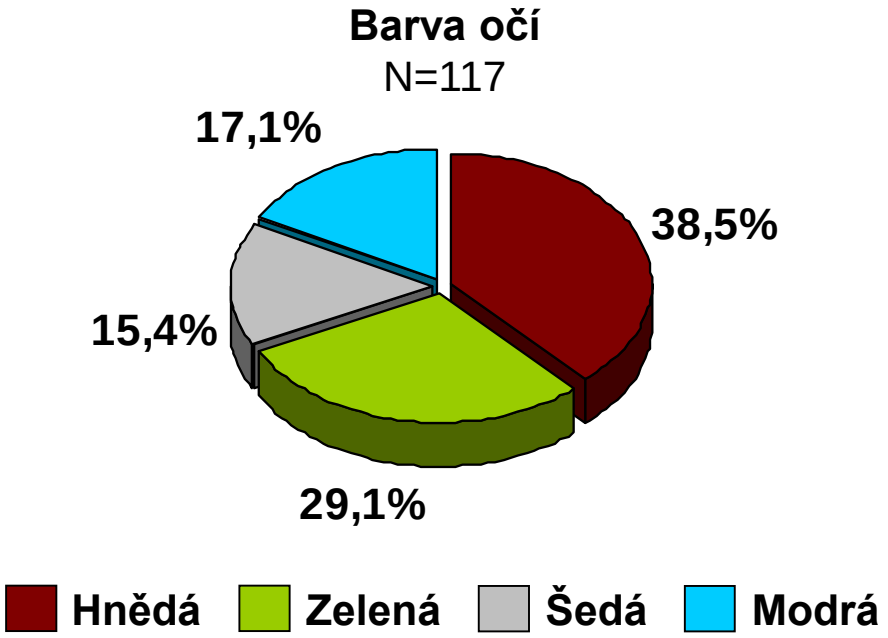


Koláčový graf je vhodné použít v prezentaci, v článku je vhodnější uvést N a %

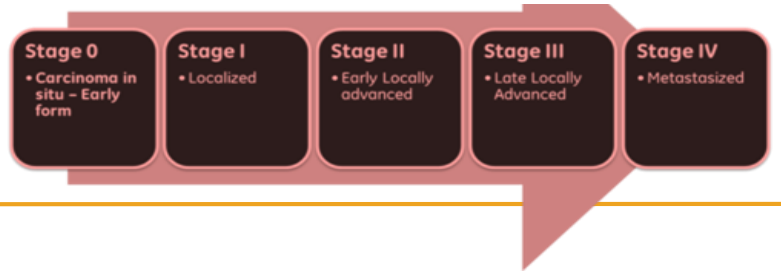
Nominální data (kvalitativní)

- Více kategorií, které nelze seřadit
- Příklady: barva očí (hnědá/zelená/...), typ skeneru (Sonata/Avanto/GE), kraj (Jihomoravský/Pardubický/...), krevní skupina (A/B/AB/0)
- Rovná se?

Koláčový graf

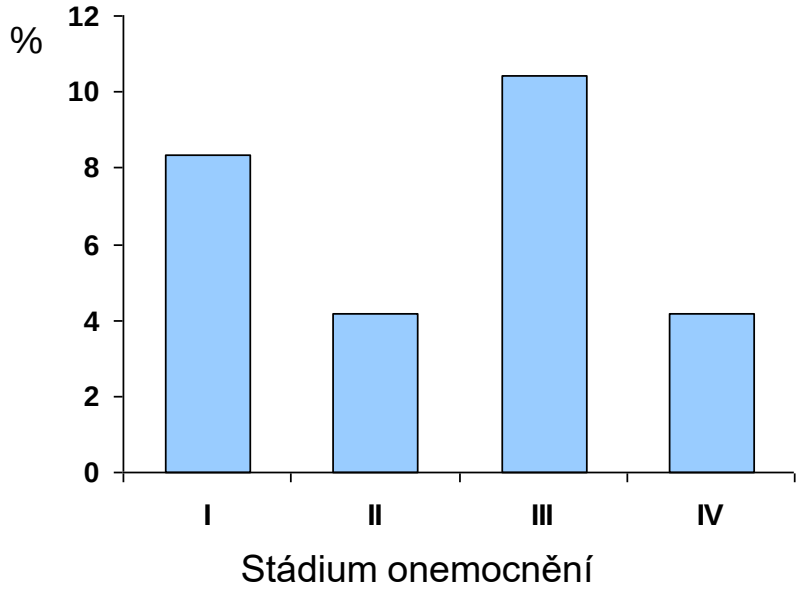


Ordinální data (kvalitativní)



- Více kategorií, které však lze seřadit
- Příklady: kategorizovaný věk (děti/lidé v produktivním věku/staří lidé), stádium onemocnění (I/II/III/IV), stupeň bolesti (mírná/střední/velká), vzdělání (ZŠ/SŠ/VŠ), četnost epileptických záchvatů (malá/střední/velká)
- Rovná se? Větší x menší?

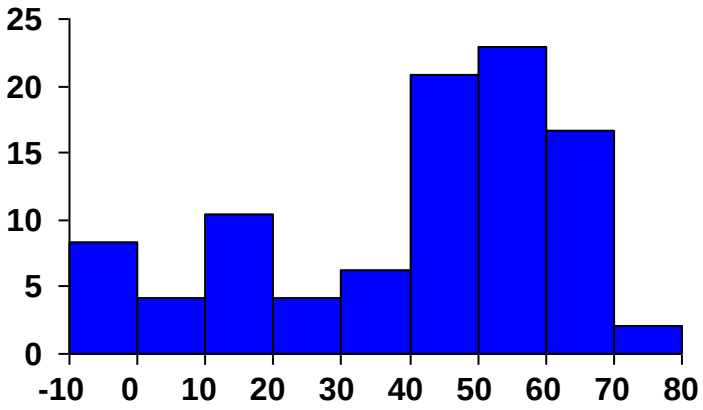
Sloupcový graf



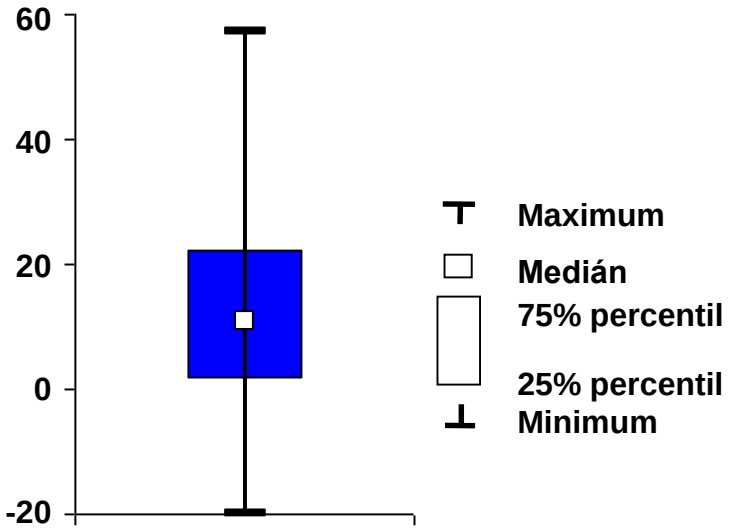
Intervalová data (kvantitativní)

- Kvantitativní data, u nichž nula byla stanovena uměle (nula nemusí vyjadřovat absenci daného znaku)
- Příklady: teplota ve stupních Celsia, kalendářní čas
- Rovná se? Větší x menší? O kolik?

Histogram



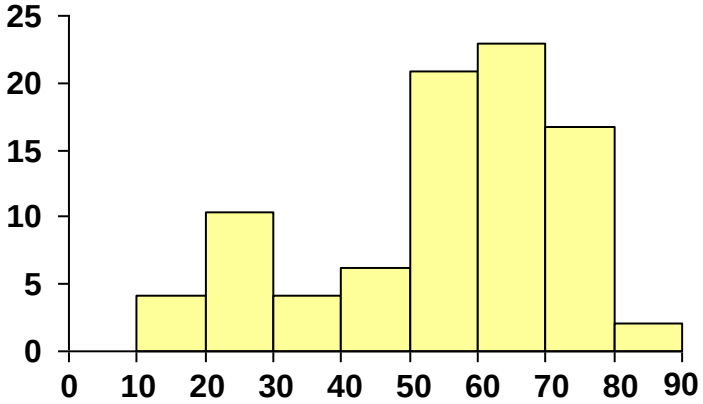
Krabicový graf (Box Plot)



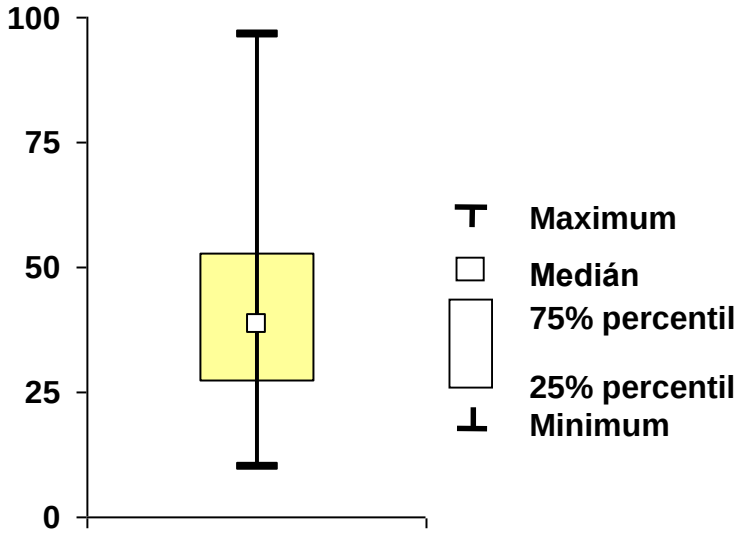
Poměrová data (kvantitativní)

- Kvantitativní data, kde nula odpovídá nepřítomnosti sledovaného znaku
- Příklady: váha, výška, objem mozkové struktury, koncentrace proteinu sAPPβ v mozkomíšním moku, počet hospitalizací pacientů
- Rovná se? Větší x menší? O kolik? Kolikrát?

Histogram

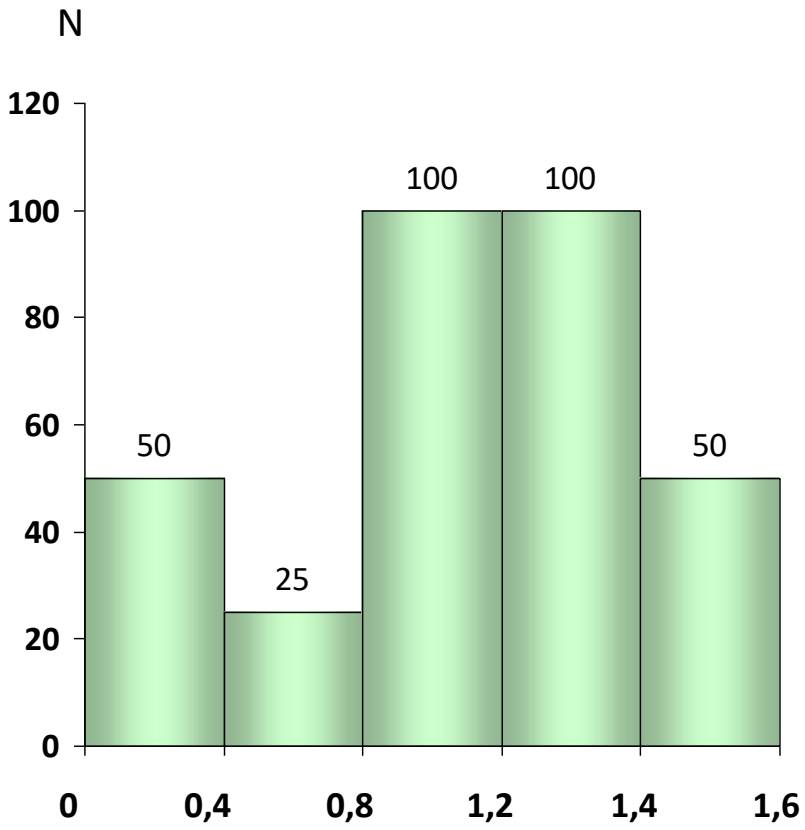


Krabicový graf (Box Plot)



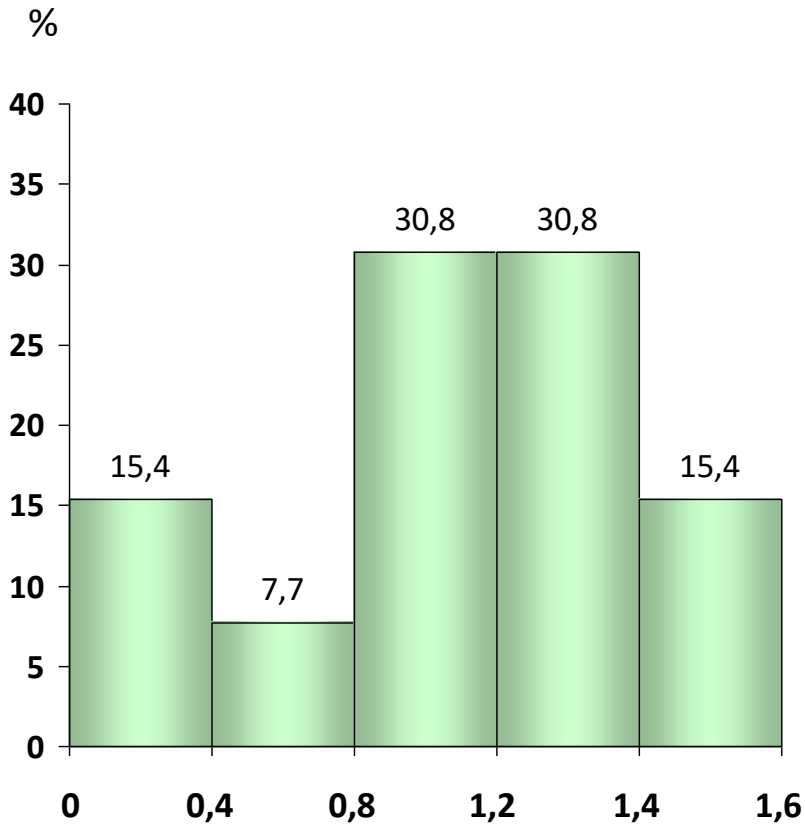
Histogramy

Histogram pro absolutní počty



→ součet je celkové N

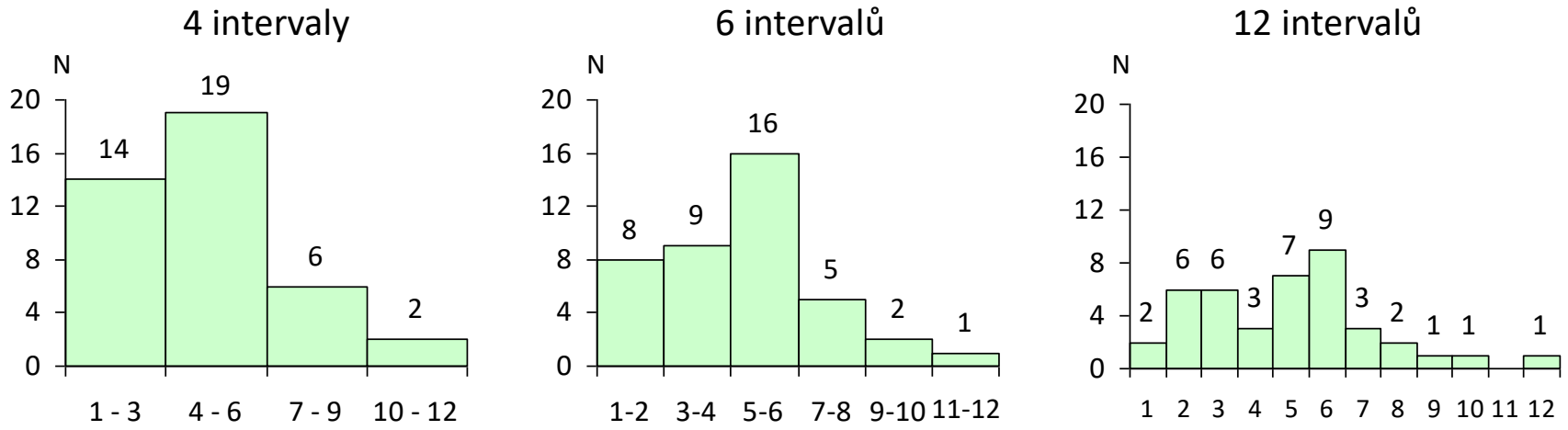
Histogram pro relativní počty



→ součet je 100%

Histogram – počet intervalů

- Počtem zvolených intervalů v histogramu rozhodujeme o tom, jak bude vypadat. Při malém počtu můžeme přehlédnout důležité prvky v datech, při velkém zase může být informace roztříštěná.



- dvě základní metody volby počtu intervalů *m*:

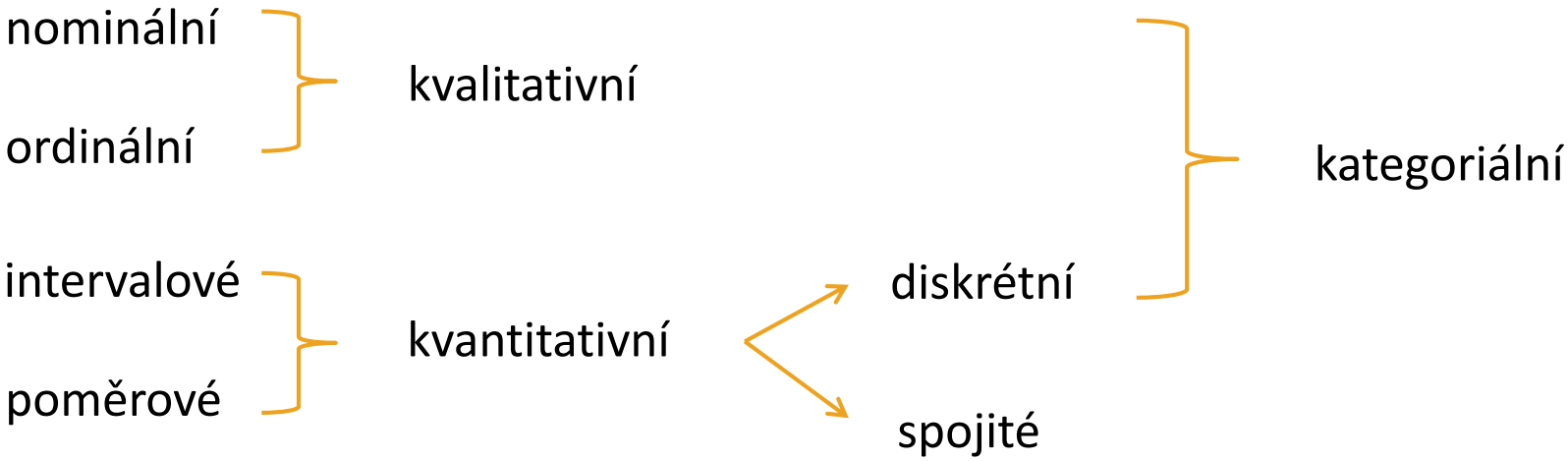
  1. odmocnina z celkového počtu: $m = \sqrt{N}$
  2. Sturgesovo pravidlo: $m = 1 + \log_2(N)$

Jiné dělení kvantitativních dat

- **Spojitá data** - mohou nabývat jakýchkoliv hodnot v určitém rozmezí
  - příklady: výška, váha, teplota, délka časového období od zahájení léčby do vymizení halucinací u schizofreniků

- **Diskrétní data** - mohou nabývat pouze spočetně mnoho hodnot
  - příklady: počet hospitalizací, počet dětí v rodině, počet krevních buněk v 1 ml krve, počet epileptických záchvatů

Shrnutí typů dat



Poznámka: diskrétní data lze zahrnovat do kategoriálních dat, pokud je počet možných hodnot proměnné malý.

Možnost převodu typu dat

Proměnné určitého typu můžeme převádět na jiný typ:

kvantitativní spojitá (věk) → ordinální (věkové kategorie) → binární (dichotomická) (<=70 let, >70 let)

Odvozené typy dat

- **Pořadí** (rank) – místo absolutních hodnot známe někdy jen jejich pořadí. Jedná se sice o ztrátu určitého množství informace, nicméně i pořadí lze v analýze využít.
- **Procento** (percentage) – sledujeme-li např. zlepšení v určitém parametru, je výhodné sledovat procentuální zlepšení. Př.: ejekční frakce levé srdeční komory.
- **Podíl** (ratio) – mnoho indexů je odvozeno jako podíl dvou měřených veličin. Př.: BMI.
- **Míra pravděpodobnosti** (rate) – týká se výskytu různých onemocnění, kdy počet nových pacientů v daném čase (studii) je vztažen na celkový počet zaznamenaných osobo-roků. Př.: výskyt nádorového onemocnění u pacientů ve studii.
- **Skóre** (score) – jedná se o uměle vytvořené hodnoty charakterizující určitý stav, který nelze jednoduše měřit jako číselné hodnoty. Př.: indexy kvality života.
- **Vizuální škála** (visual scale) – pacienti často hodnotí svoje obtíže na škále, která má formu úsečky o délce např. 10 cm. Př.: hodnocení kvality života.

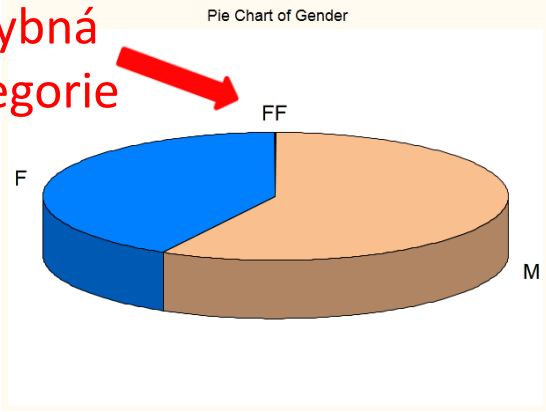
Úkol 1

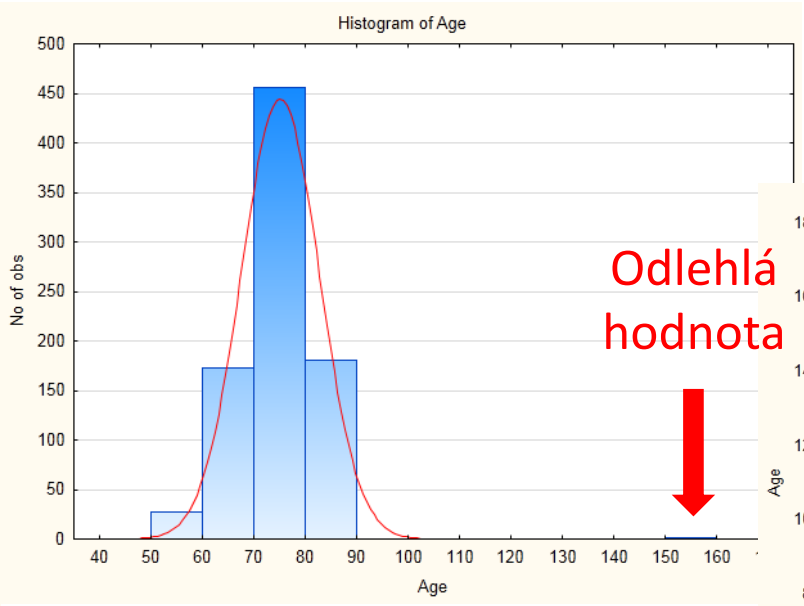
- Vykreslete koláčový graf pro typ skeneru.
- Vykreslete histogram pro objem hipokampu.
- Vykreslete krabicový graf pro objem amygdaly.

2. Předzpracování dat

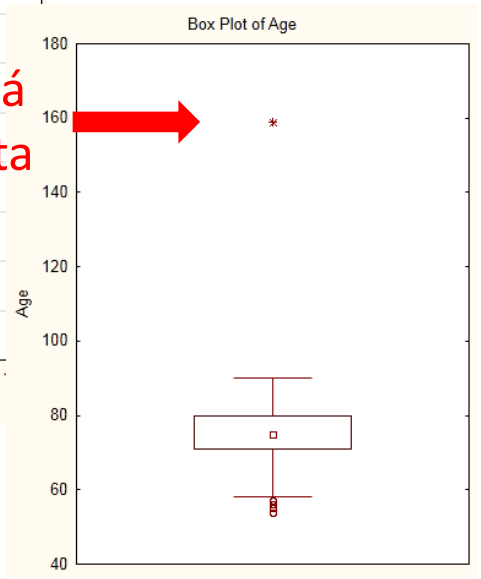
Příprava dat pro analýzu – problémy

Chybná kategorie





Odlehlá hodnota



Duplikace

	A	B	C	D	E	F	G	H	I
1	ID	Group	Gender	Age	Weight	MMSE	MMSE_24	CDR	ADAS01
13	ADNI_005_S_0553	1	M	84	66.22	30	30	0	2.33
14	ADNI_005_S_0553	1	M	84	66.22	30	30	0	2.33
15	ADNI_005_S_0602	1	M	70	85.73	29	30	0	4
16	ADNI_005_S_0610	1	M	79	88.45	29	30	0	3
17	ADNI_006_S_0484	1	M	71	91.81	29		0	2.33

Chybějící hodnota

Předzpracování dat – chybějící hodnoty

- snaha, aby v datech vůbec nenastaly
- pokud však nastanou, je silně nedoporučováno dělat každou analýzu na jinak velkém souboru (tzv. „pairwise“ odstraňování objektů) → 3 možná řešení:

1. vyloučit z analýzy všechny objekty, u nichž se vyskytla nějaká chybějící hodnota (tzv. „casewise“ = „listwise“ odstranění objektů):
   - pokud chybějících hodnot mnoho, zbyde pouze málo objektů
   - pozor na systematicky chybějící hodnoty – může dojít ke zkreslení výsledků analýz
   - občas vhodné odstranit proměnné s mnoha chybějícími hodnotami místo objektů, pokud proměnné nejsou důležité pro analýzu
2. definování souboru s vyplněnými „klíčovými“ proměnnými:
   - na tomto souboru provedena většina analýz
   - další analýzy dělány na podsouboru s menším počtem subjektů
3. doplnění chybějících hodnot (tzv. imputace):
   - doplnění průměrem z hodnot, které jsou pro danou proměnnou k dispozici
   - doplnění hodnot na základě regresních modelů
   - pozor! doplnění hodnot však může zkreslit výsledky analýz

Předzpracování dat – odlehlé hodnoty

- k identifikaci odlehlých hodnot mohou pomoci tečkové, maticové či krabicové grafy
- je třeba rozlišovat:
  1. **odlehlé hodnoty, které jsou způsobeny chybou** (měřících přístrojů apod.) - jsou to většinou nereálné hodnoty → je vhodné je smazat a dále s nimi zacházet jako s chybějícími hodnotami
  2. **odlehlé hodnoty, které jsou fyziologické** (tzn. jsou to reálné hodnoty) → je vhodné tyto hodnoty v datech ponechat, pokud je to možné a nezkreslí to analýzu a použít neparametrické metody analýzy dat
     - příklad, kdy je vhodné odlehlou hodnotu v souboru ponechat: pacienti Alzheimerovou chorobou v našem souboru mají hodnotu MMSE skóre větší než 15, jeden pacient má však hodnotu skóre 7 (je to reálná hodnota, smazáním bychom uměle snížili variabilitu)
     - příklad, kdy je nevhodné odlehlou hodnotu v souboru ponechat: chceme měřit výšku 15-letých dětí – dítě trpící nanismem měřící 80 cm by průměrnou výšku velice zkreslilo, proto ho ze souboru vyřadíme

3. Popisná sumarizace dat

Cíle popisné sumarizace dat

- zpřehlednění pozorovaných dat – ve vhodných tabulkách (a grafech)
- shrnutí pozorovaných dat (nejedná se zatím o testování)
- podklad pro stanovení hypotéz, pokud hypotézy již nejsou dány předem

- odhalení odlehlých a chybných hodnot
- odhalení chybějících hodnot (missing values)

- **sumarizace kvalitativních dat** -> cílem popsat absolutní a relativní četnosti jednotlivých kategorií
- **sumarizace kvantitativních dat** -> cílem popsat těžiště (míry polohy) a rozsah (míry variability) pozorovaných hodnot

Popisná sumarizace kvalitativních dat

Primární data

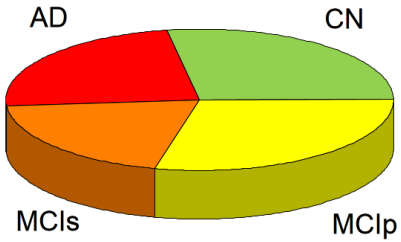
Group
AD
CN
CN
MCIp
AD
CN
MCIs
MCIp
.
.
.
.
.
.
N=833

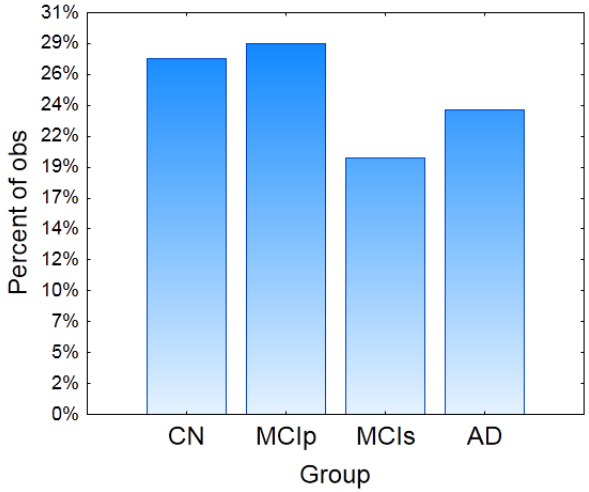
Frekvenční tabulka

x	n	%
CN	230	27,6
MCIp	240	28,8
MCIs	166	19,9
AD	197	23,6

n – absolutní četnost dané kategorie
% – relativní četnost; výpočet jako n/N

Vizualizace





K popisu lze použít i **modus** (nejčetnější pozorovaná hodnota), u ordinálních dat případně i **medián** (pokud to dává smysl).

Popisná sumarizace kvantitativních dat

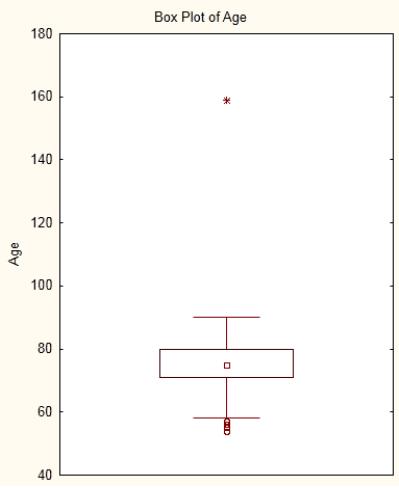
Primární data

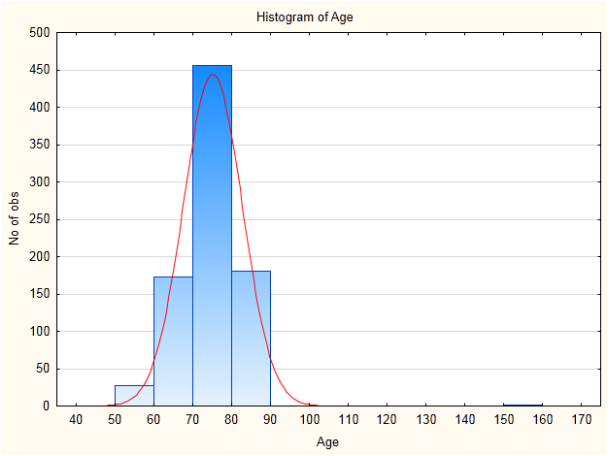
Age
84
76
79
89
71
70
88
86
.
.
.
.
.
.
N=836

Tabulka popisných statistik

	Age
N	836
Průměr (Mean)	75,0
Medián (Median)	75,0
Minimum	54,0
Maximum	159,0
Dolní kvartil (Lower Quartile)	71,0
Horní kvartil (Upper Quartile)	80,0
Směrodatná odchylka (Standard Deviation)	7,5
Variační koeficient (Coefficient of variation)	10,0

Vizualizace





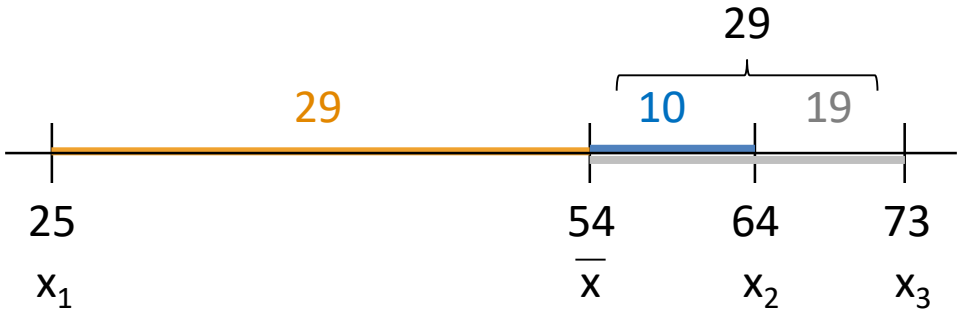
Kvantitativní data – míry polohy I

- **Minimum a maximum** – nejmenší a největší pozorovaná hodnota nám dávají obraz o tom, kde se na ose x pohybujeme.

- **Průměr** (*mean*) – charakterizuje hodnotu, kolem které kolísají ostatní pozorované hodnoty. Je to „těžiště“ dat (součet rozdílů podprůměrných hodnot od průměru je stejný jako součet rozdílů nadprůměrných hodnot od průměru).

Vizualizace:

$\bar{x}$ = (25+64+73) / 3 = 54



Příklad: N = 8

Data = 6 1 7 4 3 2 7 8

Součet dat = 6+1+7+4+3+2+7+8 = 38

Průměr = 38 / N = 38 / 8 = 4,75

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Kvantitativní data – míry polohy II

- **Medián** (*median*) – je prostřední pozorovaná hodnota. Dělí pozorované hodnoty na dvě půlky, půlka hodnot je menší a půlka hodnot je větší než medián.

- **Příklad 1:** N = 9

  N liché -> (n + 1) / 2 pozice znamená 5. pozice po seřazení

  Data = 3,0 4,2 1,1 2,5 2,2 3,8 5,6 2,7 1,7

  Seřazená data = 1,1 1,7 2,2 2,5 2,7 3,0 3,8 4,2 5,6

  Medián = 2,7

- **Příklad 2:** N = 8

  N sudé -> vypočítáme hodnotu „mezi“ 4. (n/2 -tým) a 5. (n/2+1 –tým) prvkem po seřazení

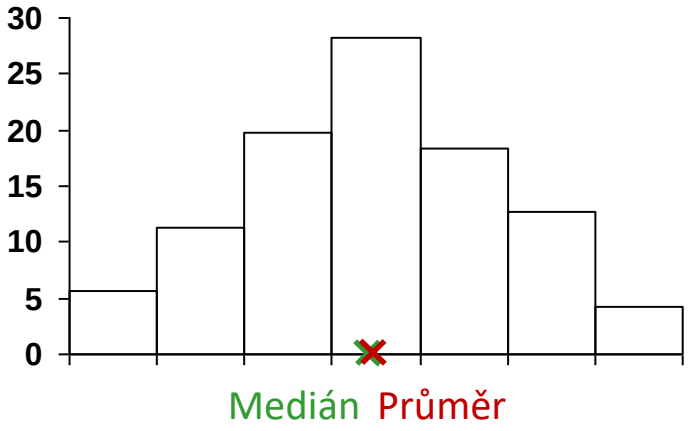
  Data = 6 1 7 4 3 2 7 8

  Seřazená data = 1 2 3 4 6 7 7 8

  Medián = (4 + 6) / 2 = 5

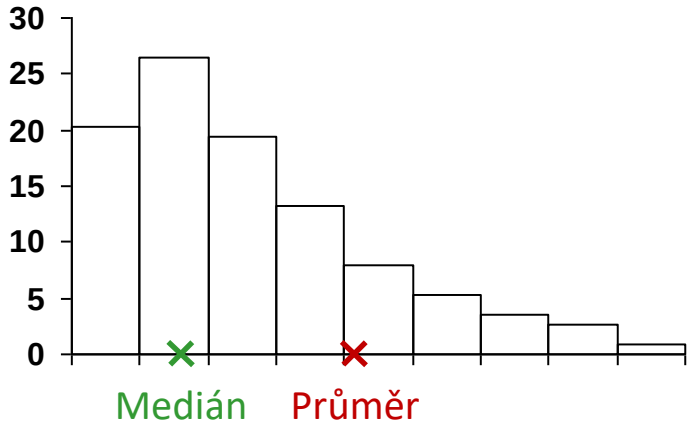
Průměr vs. medián

Symetrická data



- hodnoty mediánu a průměru téměř splývají
- medián i průměr dobrým odhadem frekvenčního středu dat (střední hodnoty)

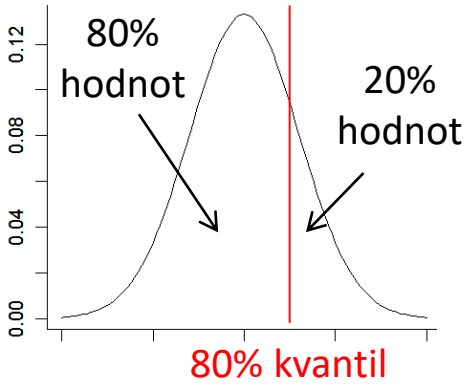
Asymetrická data



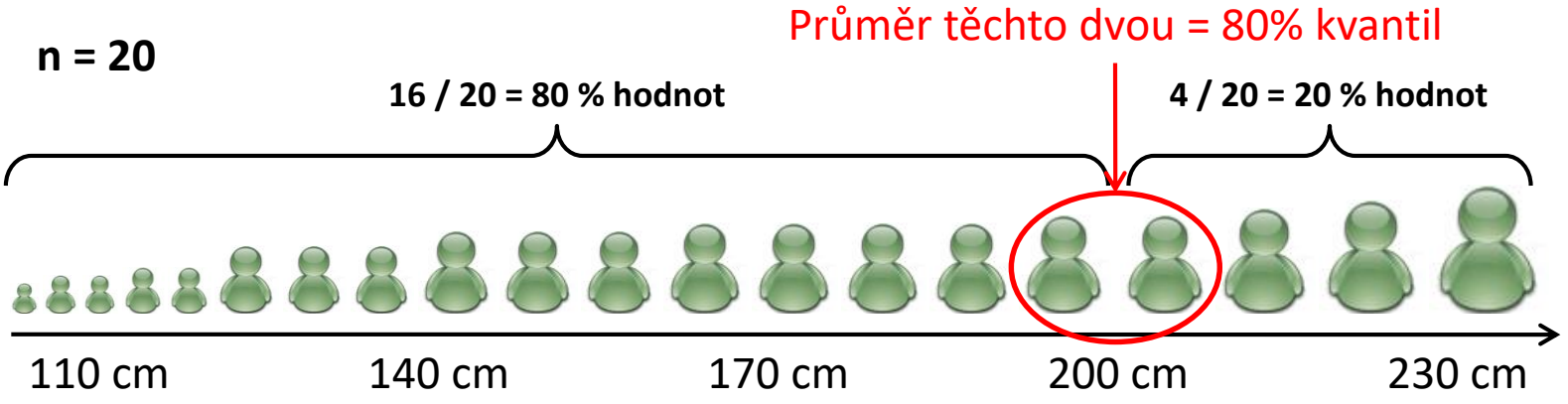
- hodnoty mediánu a průměru se liší
- průměr není vhodným odhadem frekvenčního středu dat (střední hodnoty)
- průměr vhodný, pokud chceme charakterizovat spotřebu (léků, peněz apod.)

Kvantil (*quantile*)

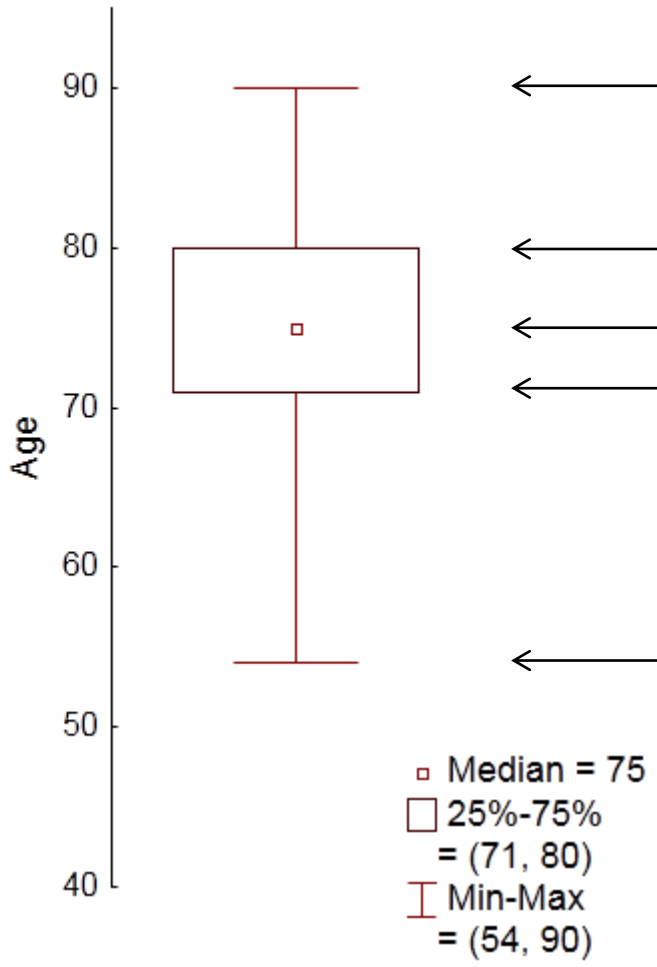
- Kvantil lze definovat jako číslo na reálné ose, které rozděluje pozorovaná data na dvě části: *p*% kvantil rozděluje data na *p* % hodnot a (100-*p*) % hodnot.



- Máme soubor 20 osob, u nichž měříme výšku. Chceme zjistit 80% kvantil souboru pozorovaných dat.

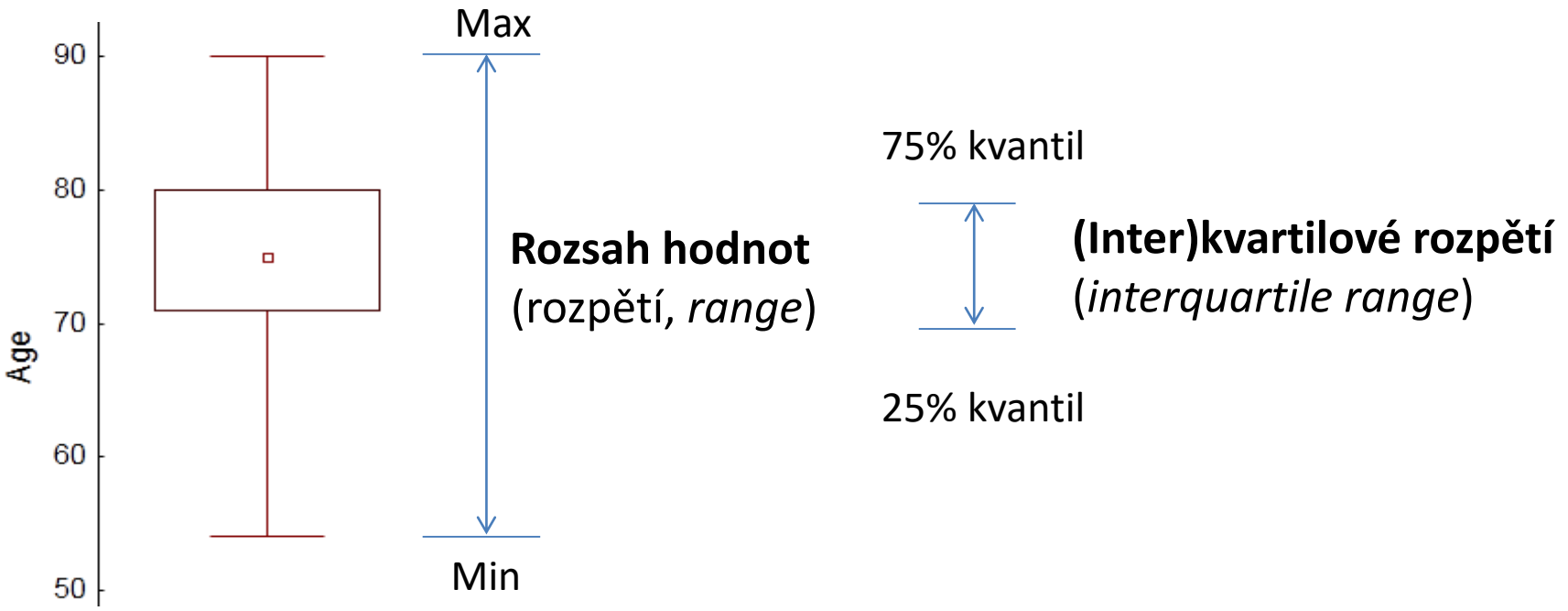


Významné kvantily



	Age
Maximum = 100% kvantil	90
Horní kvartil = 75% kvantil	80
Medián = 50% kvantil	**75**
Dolní kvartil = 25% kvantil	71
Minimum = 0% kvantil	54

Kvantitativní data – míry variability I



- **Rozsah hodnot** (rozpětí, *range*) = maximum – minimum. Je to nejjednodušší charakteristika variability pozorovaných dat. Je snadno ovlivnitelný netypickými (odlehlými) hodnotami.
- **Kvantilové rozpětí** je definováno p% kvantilem a (100-p)% kvantilem a je méně ovlivněno odlehlými hodnotami. Speciálním případem je (inter)kvartilové rozpětí (= 75% kvantil – 25% kvantil), které pokrývá 50% pozorovaných hodnot.

Kvantitativní data – míry variability II

- **Rozptyl** – průměrný čtverec odchylky od průměru. Velmi ovlivnitelný odlehlými hodnotami.

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

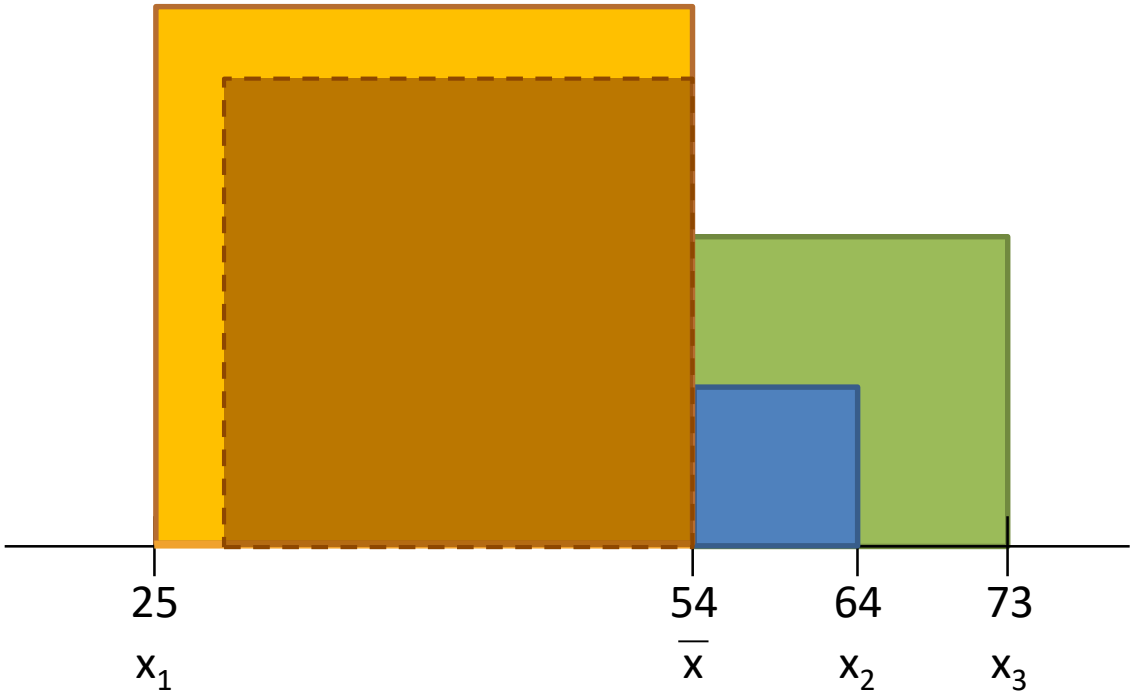
- **Směrodatná odchylka** – odmocnina z rozptylu. Výhodou směrodatné odchylky je, že má stejné jednotky jako pozorovaná data.

- **Variační koeficient (koeficient variace)** – podíl směrodatné odchylky a průměru. Používá se na srovnání variability mezi datovými soubory. Často se vyjadřuje v procentech.

$$v = \frac{s}{\bar{x}} \cdot 100\ \%$$

Výpočet rozptylu a směrodatné odchylky - ukázka

- Příklad čtverců odchylek od průměru pro *n* = 3.
- Rozptyl je možno značně ovlivnit odlehlými pozorováními.



$$s = \sqrt{\frac{(25-54)^2 + (64-54)^2 + (73-54)^2}{2}} = \sqrt{651} = 25{,}5$$

Rozptyl:

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

Směrodatná odchylka:

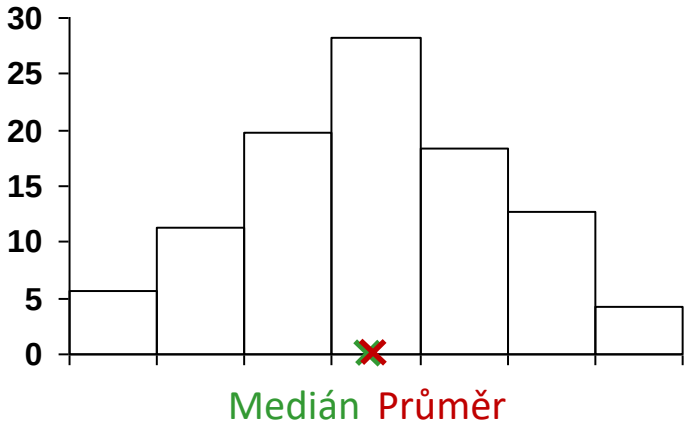
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Úkol 2

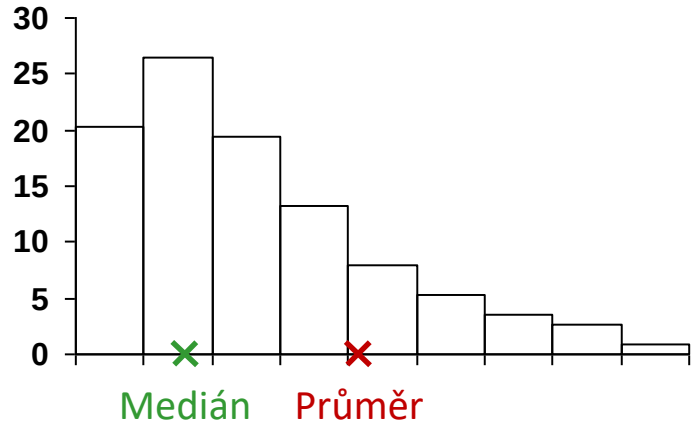
- Proveďte popisnou sumarizaci pohlaví.
- Proveďte popisnou sumarizaci objemu všech šesti mozkových struktur (do jedné tabulky).

Popis kvantitativních dat – shrnutí

Symetrická data



Asymetrická data



	Age
N	833
Průměr (Mean)	74,8
Směrodatná odchylka (SD)	6,9
95% interval spolehlivosti (CI)	74,3-75,3
Medián (Median)	75,0
Minimum	54,0
Maximum	90,0

	MMSE
N	833
Medián (Median)	27
Minimum	18
Maximum	30

Poděkování…

Příprava výukových materiálů předmětu „DSAN01 Analýza dat pro Neurovědy “ byla finančně podporována prostředky projektu FRVŠ č. 942/2013 „Inovace materiálů pro interaktivní výuku a samostudium předmětu Analýza dat pro Neurovědy“