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, jejich vizualizace a popisná sumarizace
   - Modelová rozdělení dat, transformace dat
   - Intervaly spolehlivosti
2. Jak medicínská data správně testovat :
   - Formulování hypotéz, hladina významnosti, síla testu, p-hodnota
   - Jednovýběrové testy: z-test, jednovýběrový t-test, párový t-test
3. Jak a kdy použít parametrické a neparametrické testy I. :
   - 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.

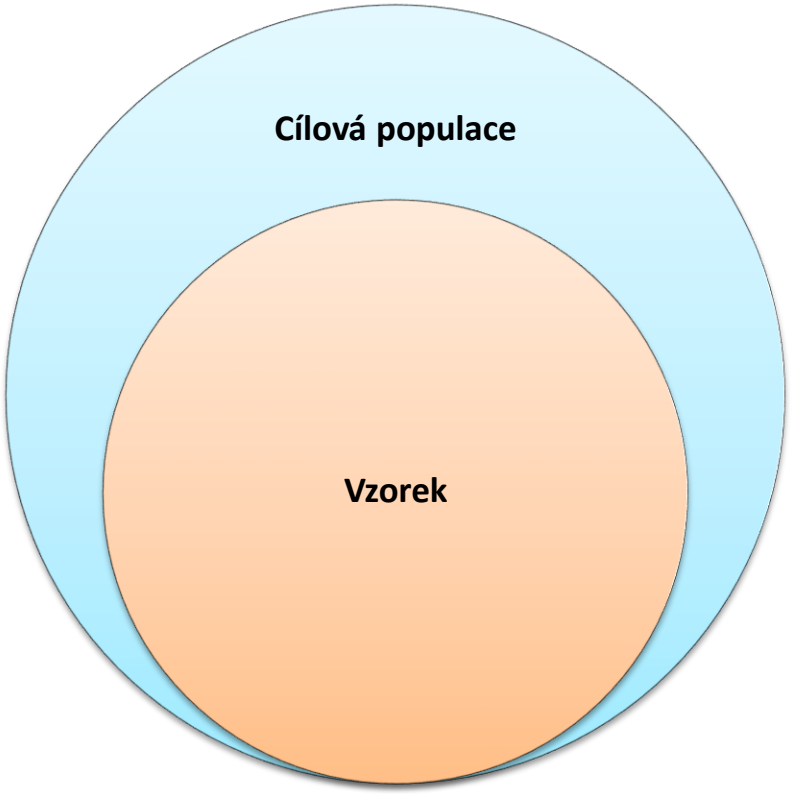
Osnova

1. Typy medicínských dat a jejich vizualizace
2. Popisná sumarizace dat
3. Normální rozdělení a rozdělení od něj odvozená
4. Transformace dat
5. Intervaly spolehlivosti

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)

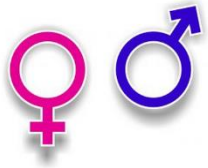
ID	Pohlaví	Věk	Váha	…
1	muž	84	85,5	
2	žena	25	62,0	
3				
4				
…				

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



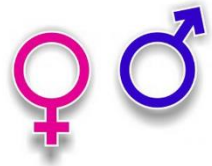
  - Nominální data

  - Ordinální data

- **Kvantitativní data:**
  - Intervalová data

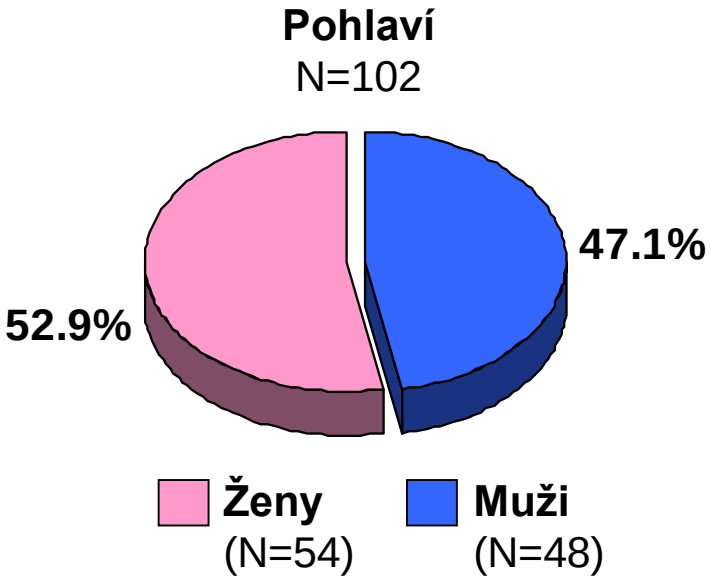
  - Poměrová data

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

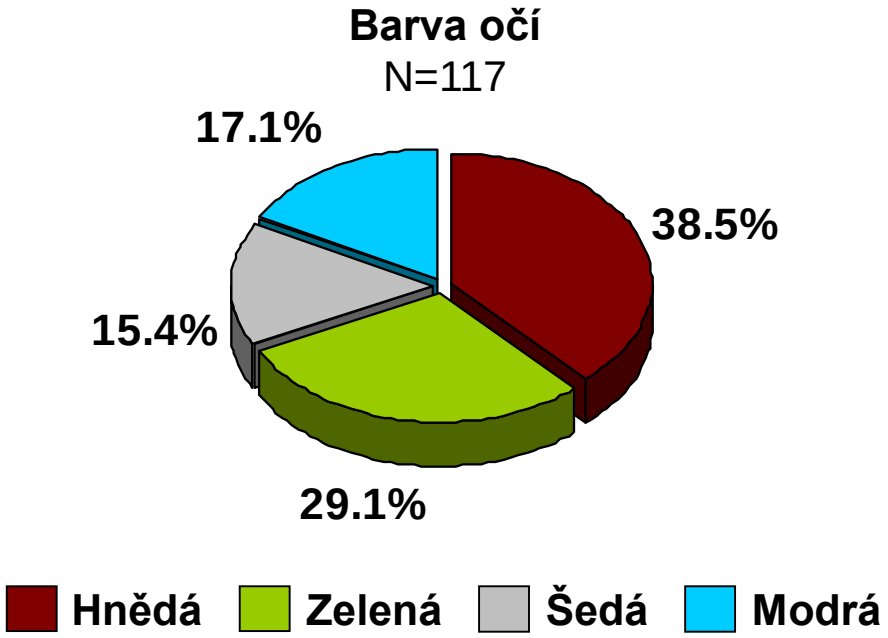


Koláčový graf je vhodné použít v prezentaci, v článku je vhodnější uvést N a %

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

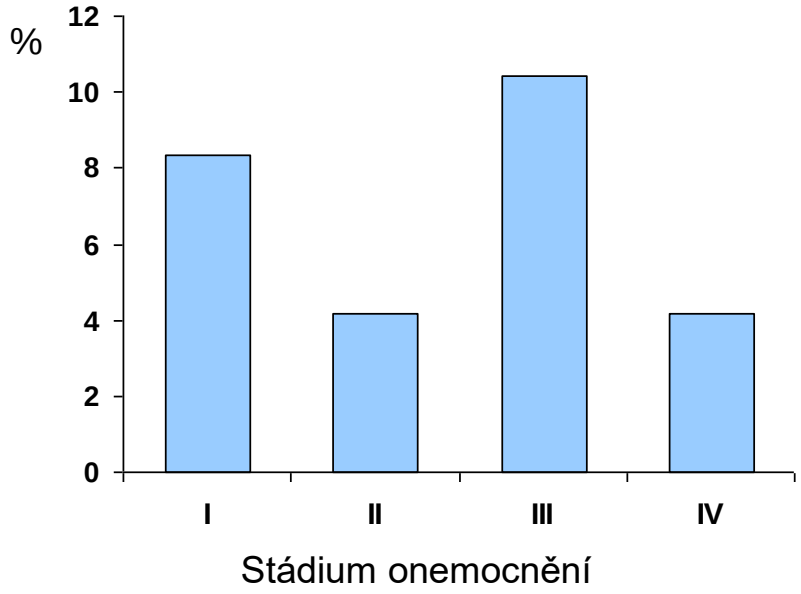


Ordinální data (kvalitativní)



- 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?

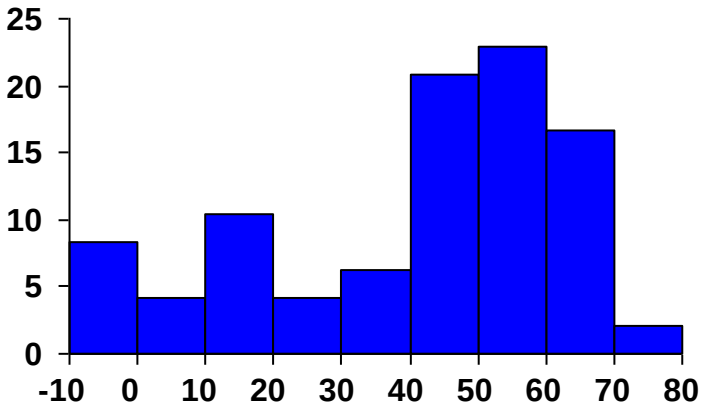
Sloupcový graf



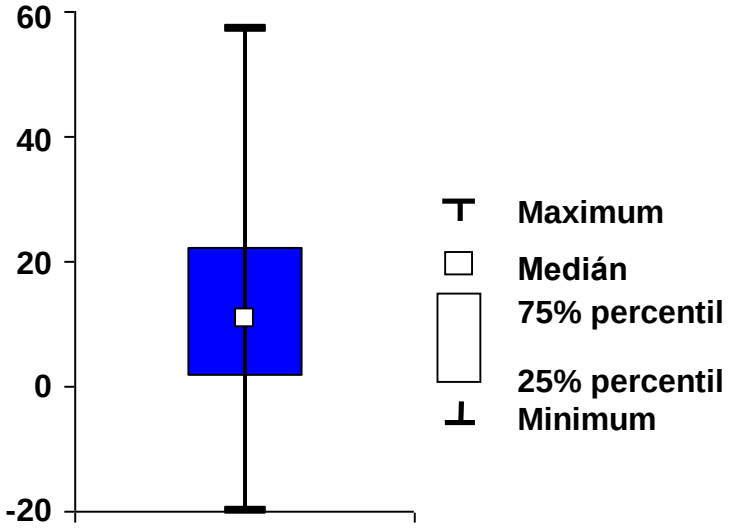
Intervalová data (kvantitativní)

- 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



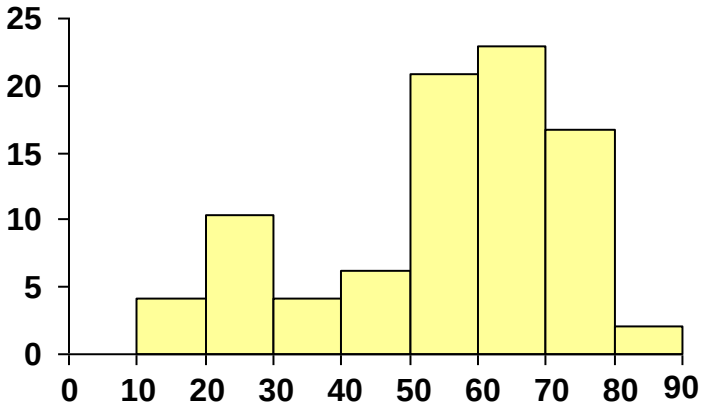
Krabicový graf (Box Plot)



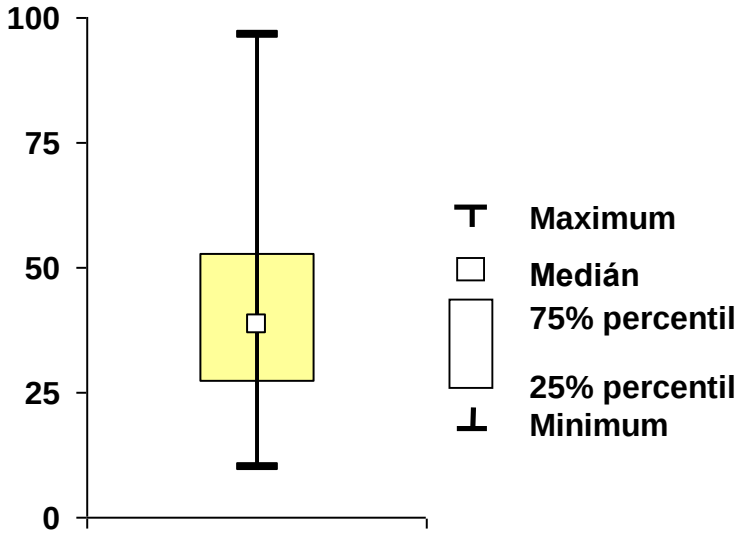
Poměrová data (kvantitativní)

- 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

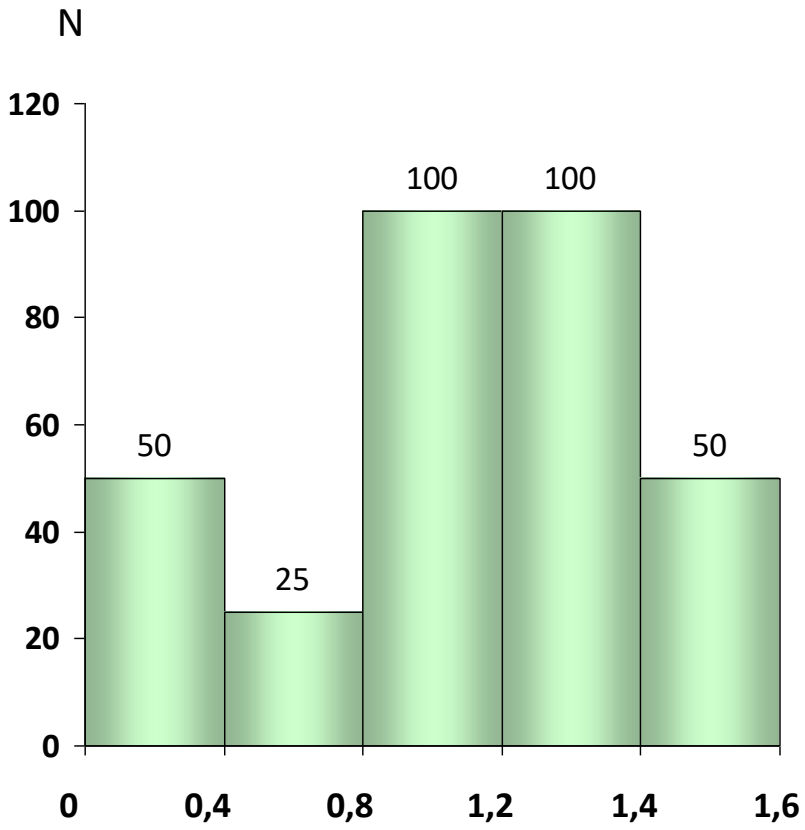


Krabicový graf (Box Plot)



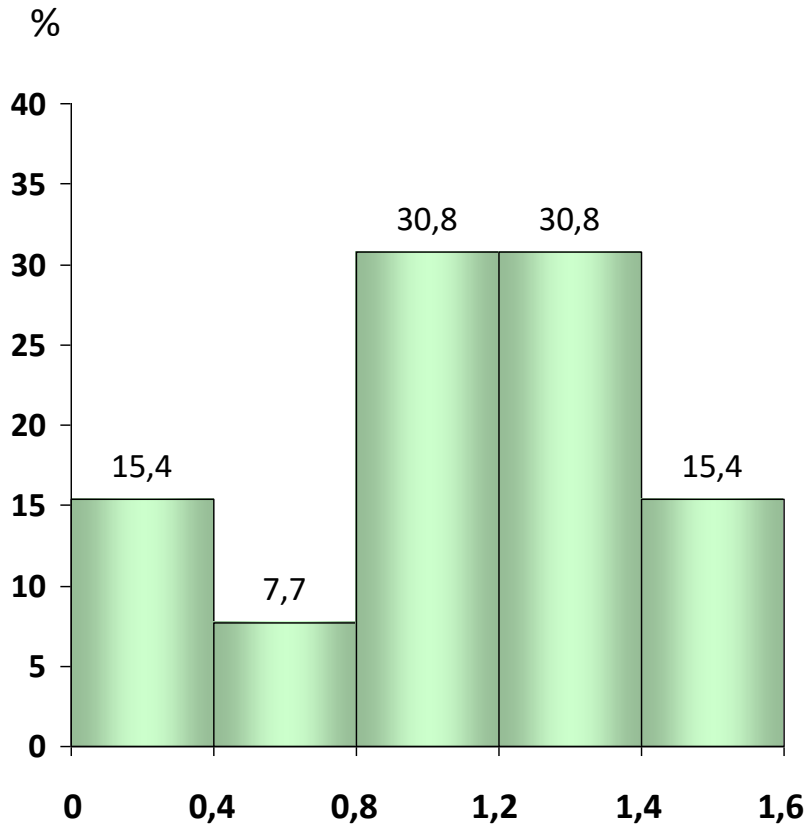
Histogramy

Histogram pro absolutní počty



→ součet je celkové N

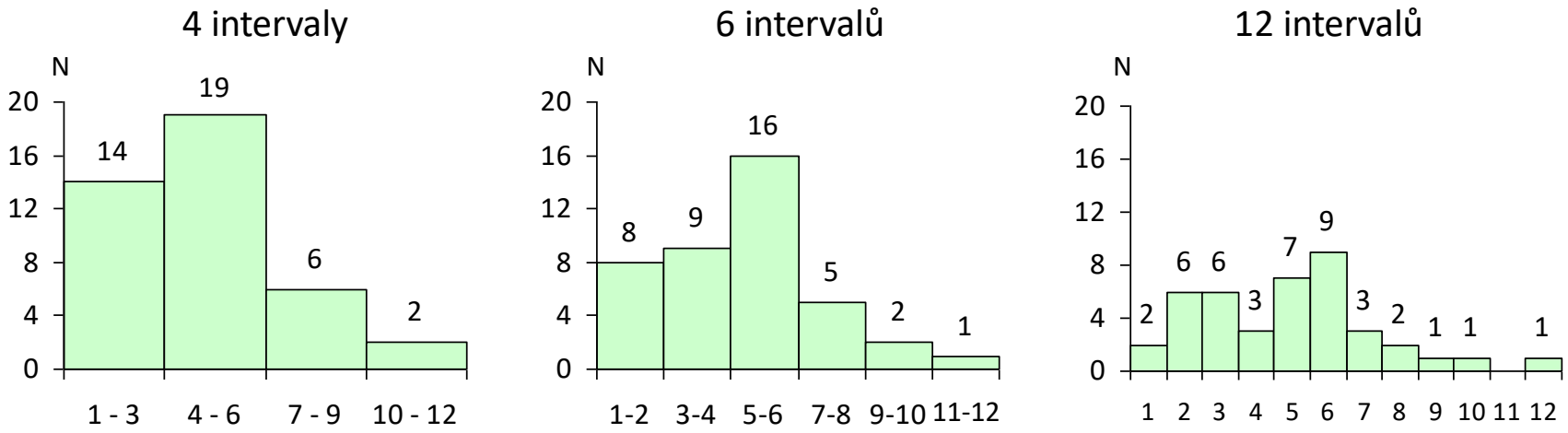
Histogram pro relativní počty



→ 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*:

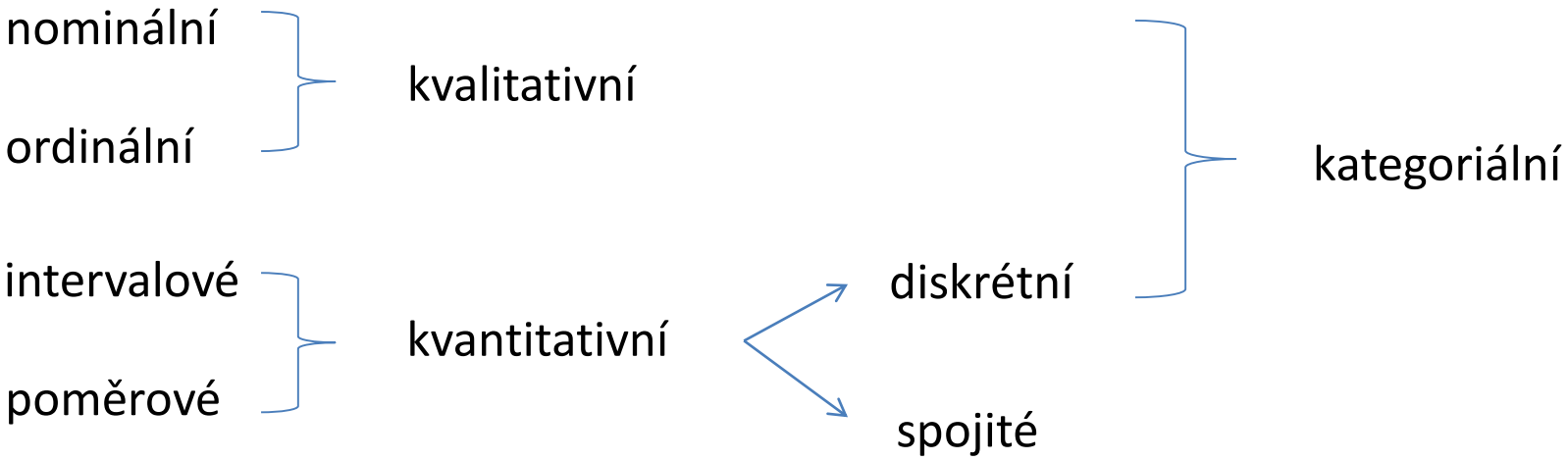
  1. odmocnina z celkového počtu: $m = \sqrt{N}$
  2. Sturgesovo pravidlo: $m = 1 + \log_2(N)$

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



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.

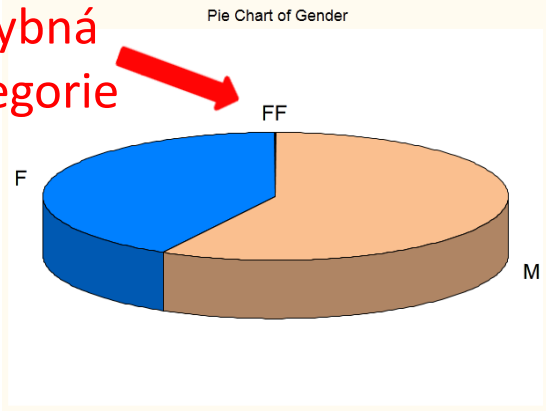
Úkol 1

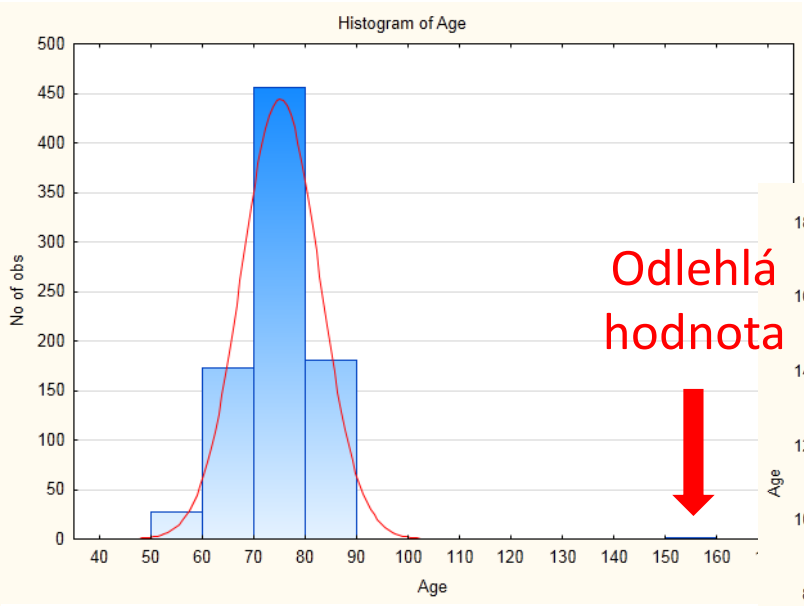
- Vykreslete koláčový graf pro typ skeneru.
- Vykreslete histogram pro objem hipokampu.
- Vykreslete krabicový graf pro objem amygdaly.

2. Popisná sumarizace dat

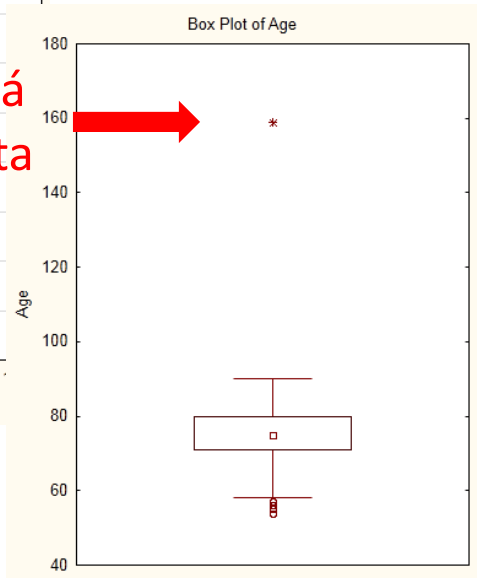
Příprava dat pro analýzu – problémy

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
- další možné metody k identifikaci odlehlých hodnot budou probrány na příští přednášce
- 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

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

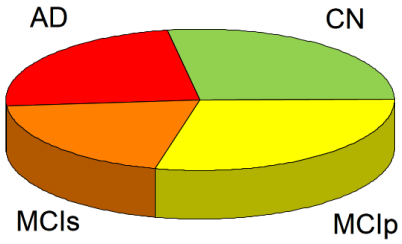
Group
AD
CN
CN
MCIp
AD
CN
MCIs
MCIp
.
.
.
.
.
.
N=833

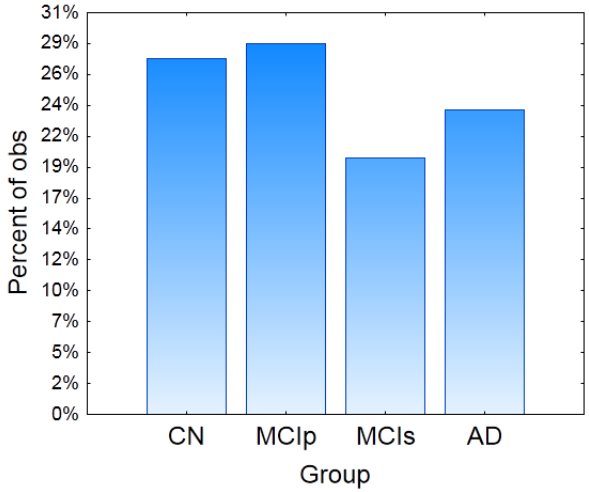
Frekvenční tabulka

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

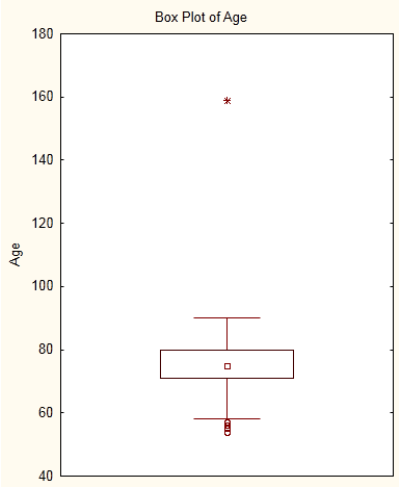
Primární data

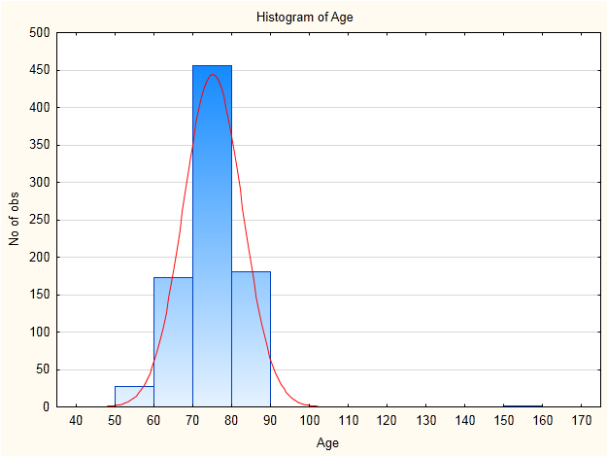
Age
84
76
79
89
71
70
88
86
.
.
.
.
.
.
N=836

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

- **Minimum a maximum** – nejmenší a největší pozorovaná hodnota nám dávají obraz o tom, kde se na ose x pohybujeme.
- **Průměr** – charakterizuje hodnotu, kolem které kolísají ostatní pozorované hodnoty. Je to „těžiště“ dat (součet podprůměrných hodnot je stejný jako součet nadprůměrných hodnot).

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

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

$$\tilde{x} = x_{((n+1)/2)} \quad \text{pro } n \text{ liché}$$

$$\tilde{x} = \tfrac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right) \quad \text{pro } n \text{ sudé}$$

Hodnoty *x* jsou seřazené podle velikosti.

Výpočet mediánu - příklady

$\tilde{x} = x_{((n+1)/2)}$ pro *n* liché

$\tilde{x} = \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)})$ pro *n* sudé

- **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

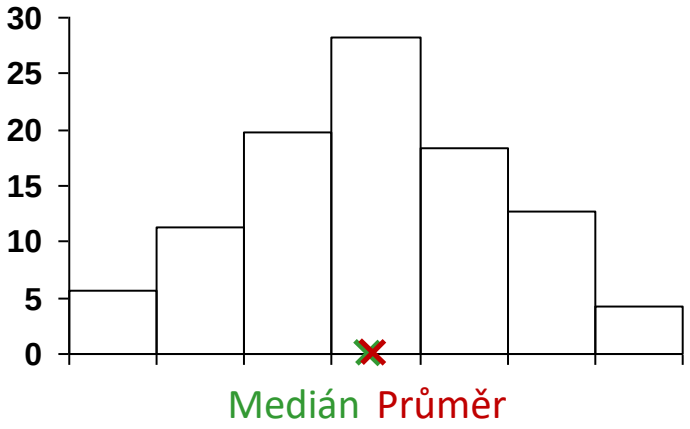
‖ Data = 6 1 7 4 3 2 7 8

‖ Seřazená data = 1 2 3 4 6 7 7 8

‖ Medián = (4 + 6) / 2 = 5

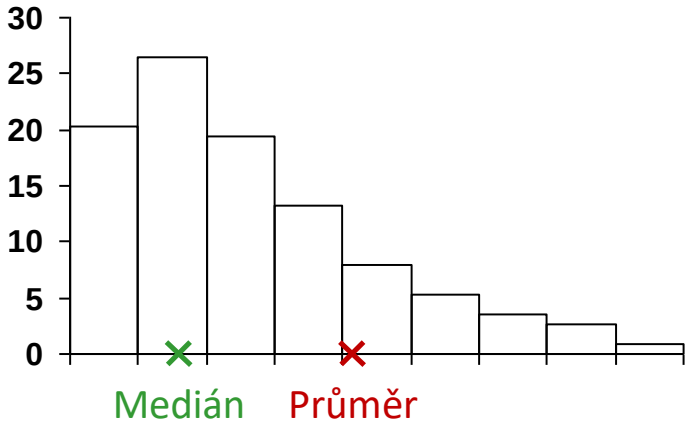
Průměr vs. medián

Symetrická data



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

Asymetrická data



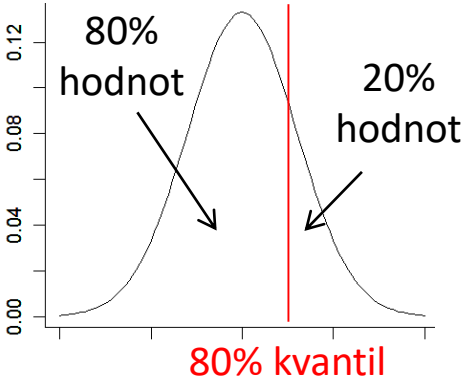
- hodnoty mediánu a průměru se liší
- průměr není vhodným odhadem frekvenčního středu dat (střední hodnoty)
- průměr vhodný, pokud chceme charakterizovat spotřebu (léků, peněz apod.)

Kvantil

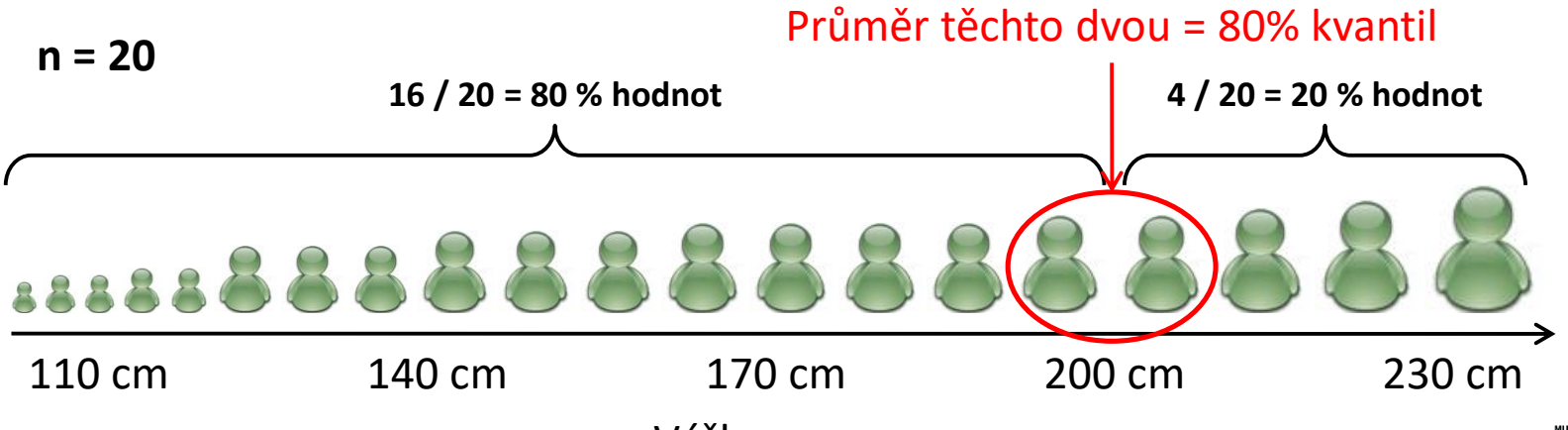
- Kvantil lze definovat jako číslo na reálné ose, které rozděluje pozorovaná data na dvě části: *p*% kvantil rozděluje data na *p* % hodnot a (100-*p*) % hodnot.

$$x_p = x_{(k+1)} \qquad \text{pro } k \neq np$$

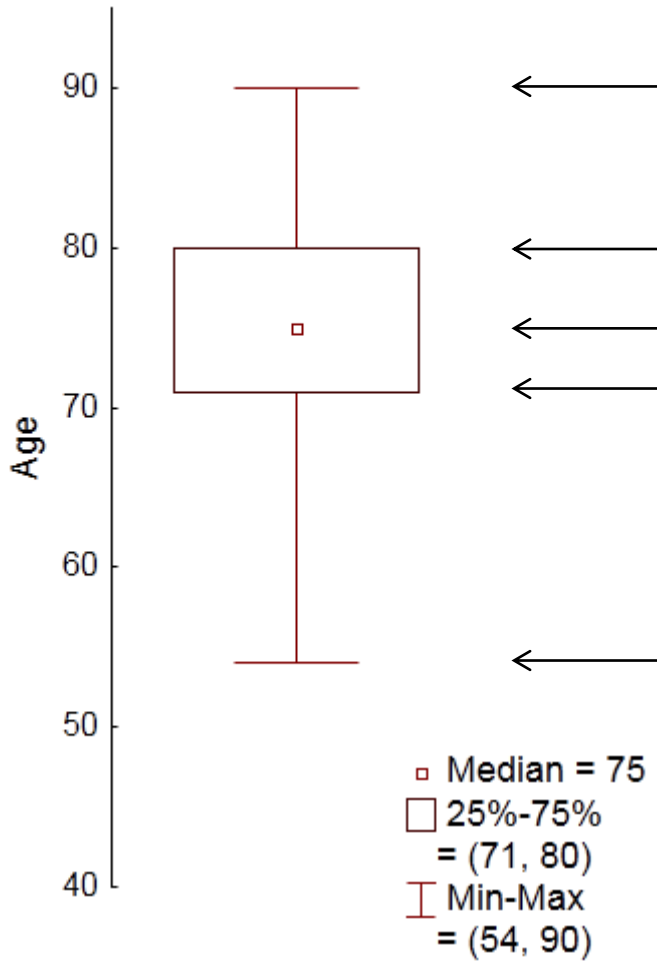
$$x_p = \tfrac{1}{2}\left(x_{(k)} + x_{(k+1)}\right) \qquad \text{pro } k = np$$



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

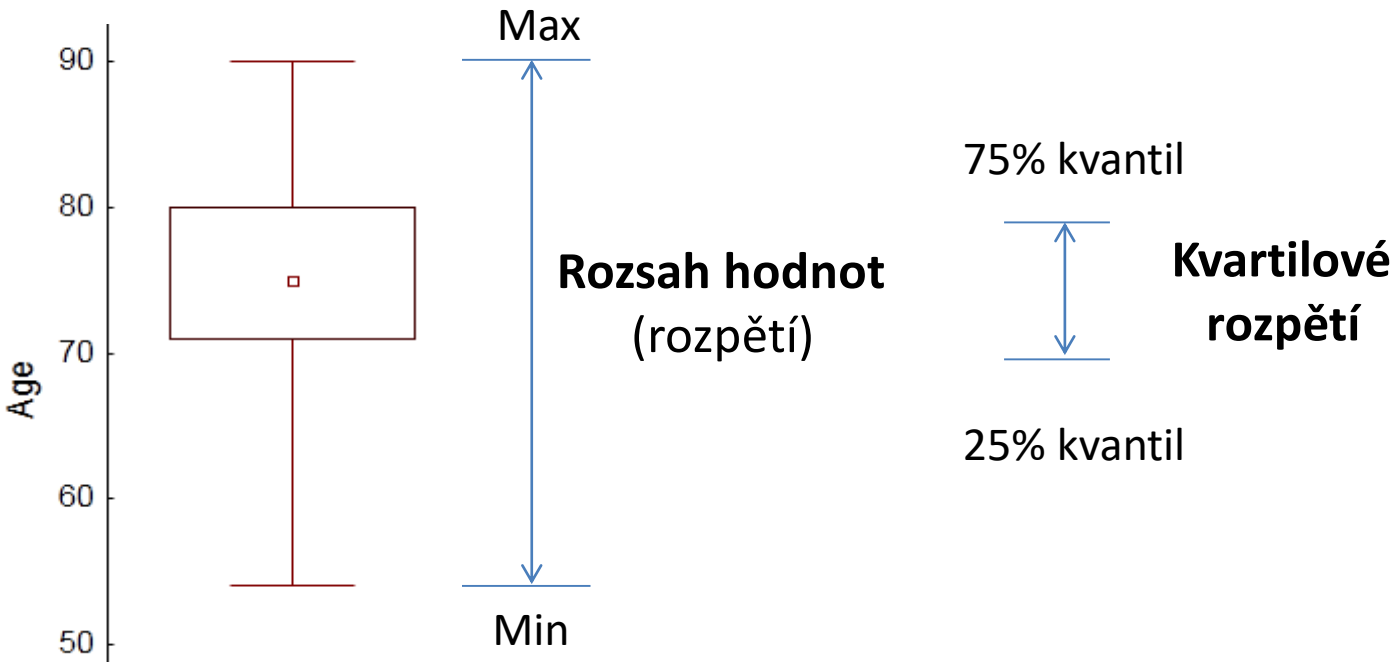


Významné kvantily



	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) = 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 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.

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

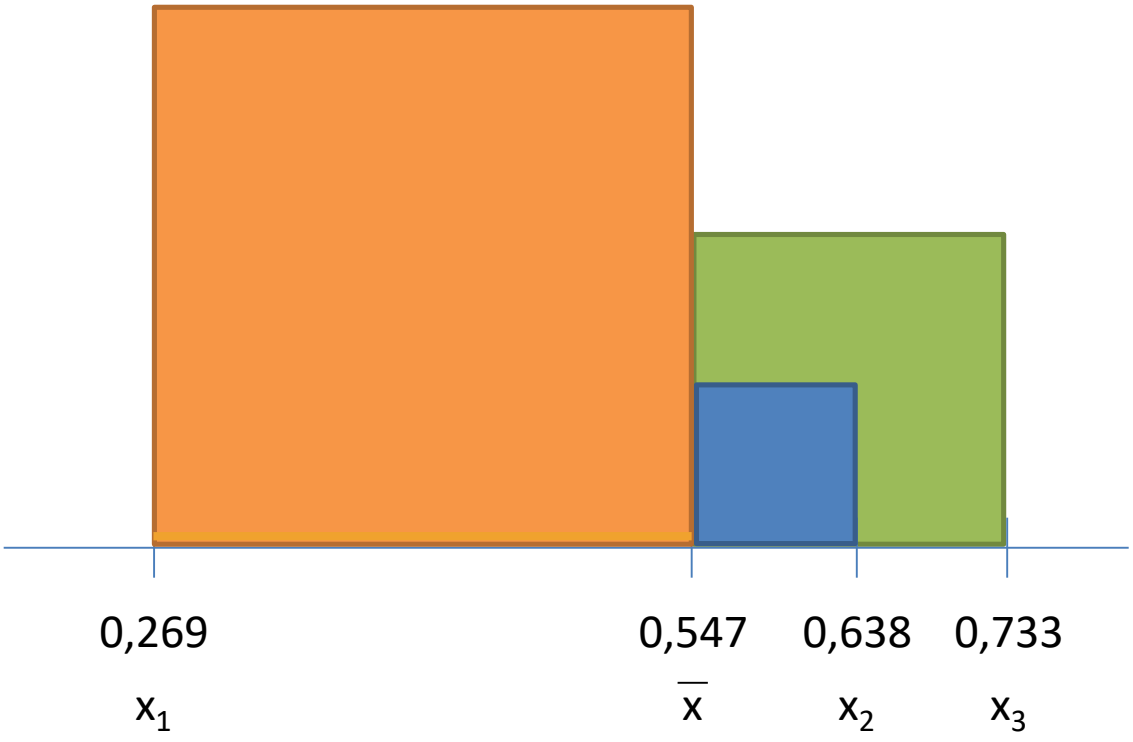
- **Směrodatná odchylka** – odmocnina z rozptylu. Výhodou směrodatné odchylky je, že má stejné jednotky jako pozorovaná data.

- **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.



Rozptyl:

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

Směrodatná odchylka:

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

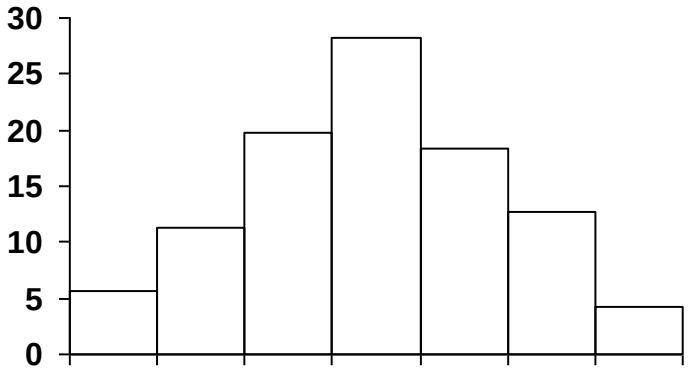
Úkol 2

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

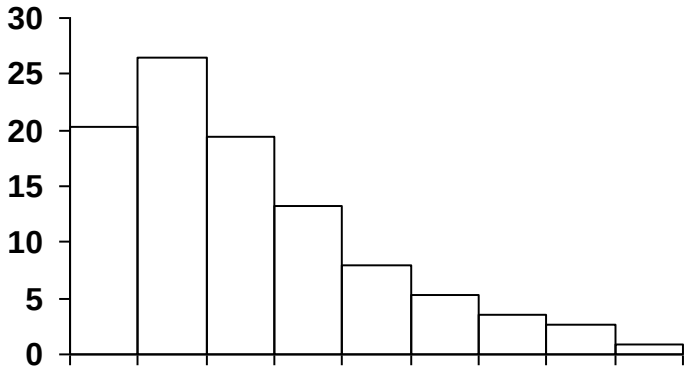
3. Vybraná modelová rozdělení

Motivace

Symetrická data



Asymetrická data



K čemu je nám znalost o modelových rozděleních?

- **Popis vlastností cílové populace** – na základě pozorovaných dat (histogram, box plot, popisné statistiky) jsme schopni usuzovat na charakter rozdělení pravděpodobnosti sledované veličiny. Dokonce jsme schopni otestovat míru shody s teoretickým rozdělením.

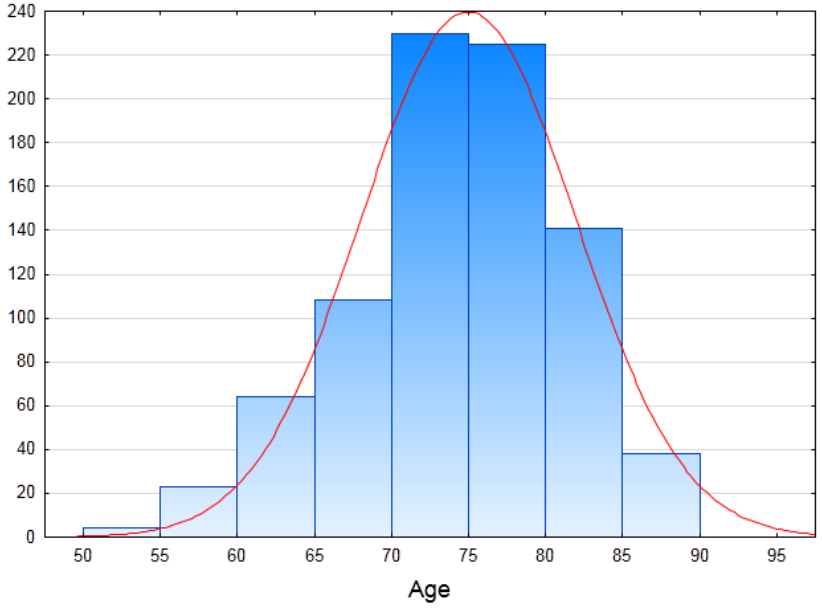
- **Srovnání vlastností cílové populace/populací** – na základě pozorovaných dat a našich předpokladů o teoretickém modelu (hypotéz) jsme schopni pomocí statistických testů srovnávat vlastnosti jedné nebo více cílových populací.

- **Predikce vlastností cílové populace** – nevyvrátíme-li na základě pozorovaných dat platnost teoretického modelu, jsme schopni se ptát, jak a s jakou pravděpodobností se bude cílová populace v budoucnu chovat.

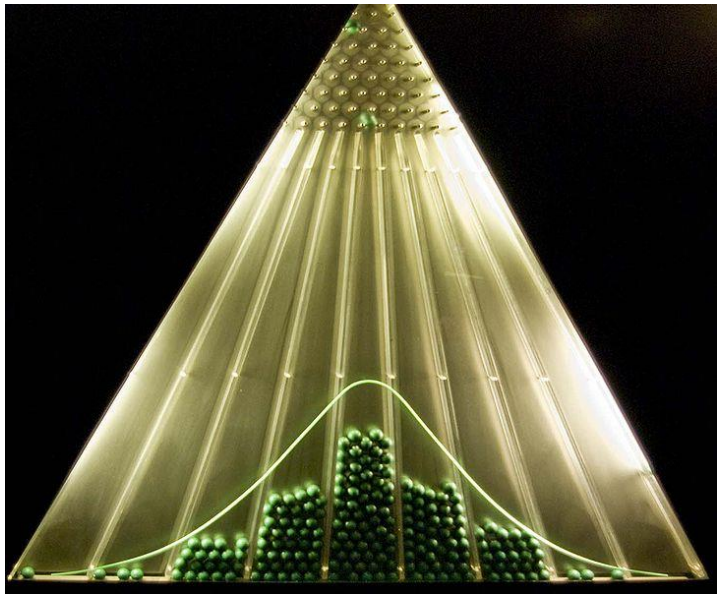
Normální rozdělení

- jiný název – Gaussovo rozdělení
- základní rozdělení – u mnoha klinických a biologických veličin: tělesná výška, délka končetin a kostí, krevní tlak,...
- hodnoty veličiny se symetricky shlukují kolem středu, variabilita je dána aditivním vlivem mnoha „slabě působících faktorů“

Příklad - věk



Příklad vzniku normálního rozdělení – Galtonova deska



Normální rozdělení

- **střední hodnota** – sumární statistika středu dat (tzn. číslo, které zastoupí střední, typickou, průměrnou hodnotu)

  - u normálního rozd. označení: **μ**

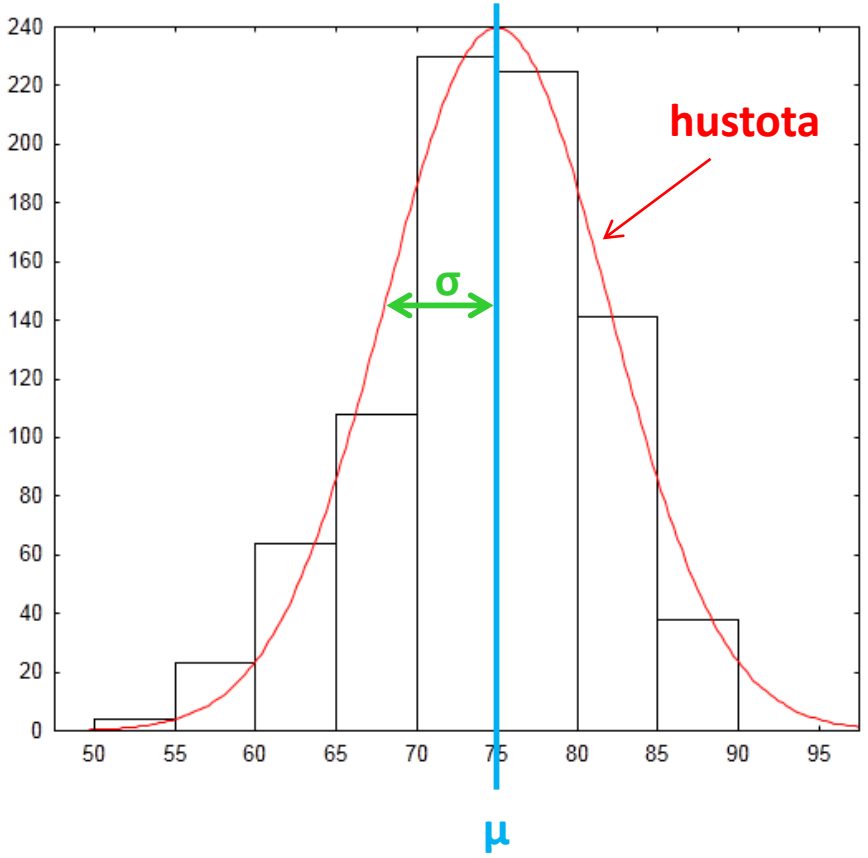
- **rozptyl** – sumarizace variability (tzn. odlišnosti jedinců zahrnutých ve výběrovém souboru);

  - u normálního rozd. označení: **σ^2**

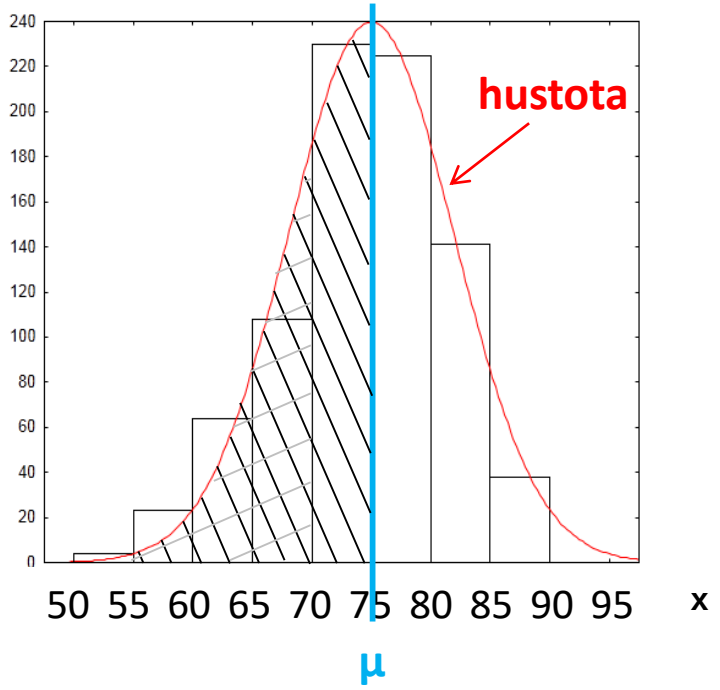
- tvar rozdělení nám popisuje **hustota** (hustota normálního rozdělení – tzv. Gaussova křivka):

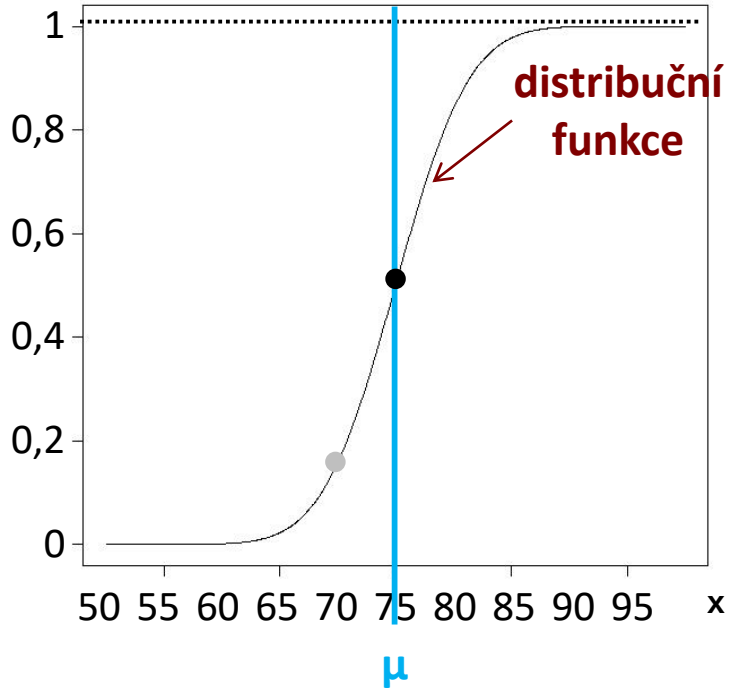
$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

- značení: ***N*(*μ*,*σ*²)**



Normální rozdělení – distribuční funkce

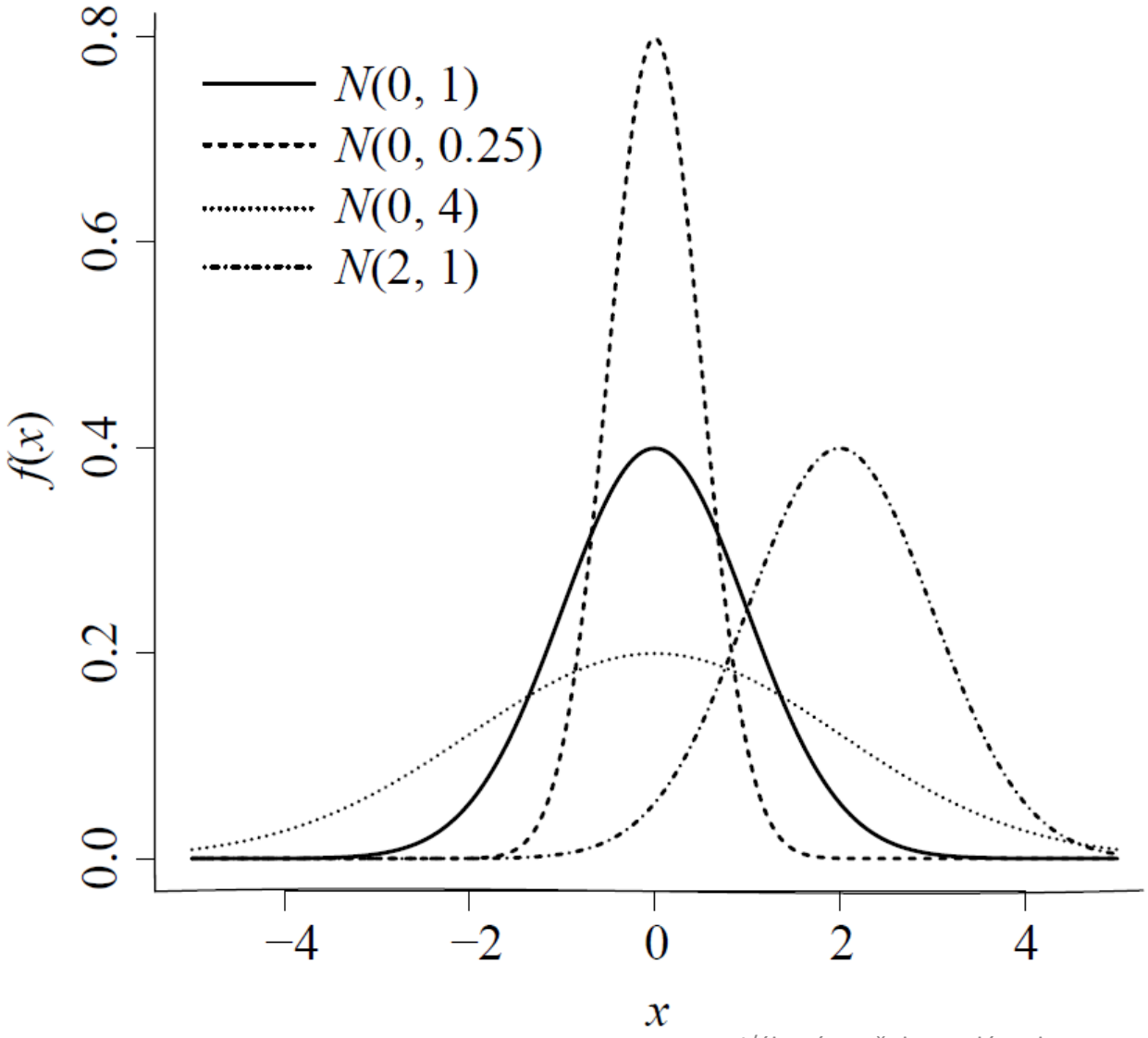




interval	d(l)	n(l)	n(l)/n	N(x'')	F(x'')
<50,55)	5	4	0,005	4	0,005
<55,60)	5	23	0,028	27	0,033
<60,65)	5	64	0,077	91	0,110
...					

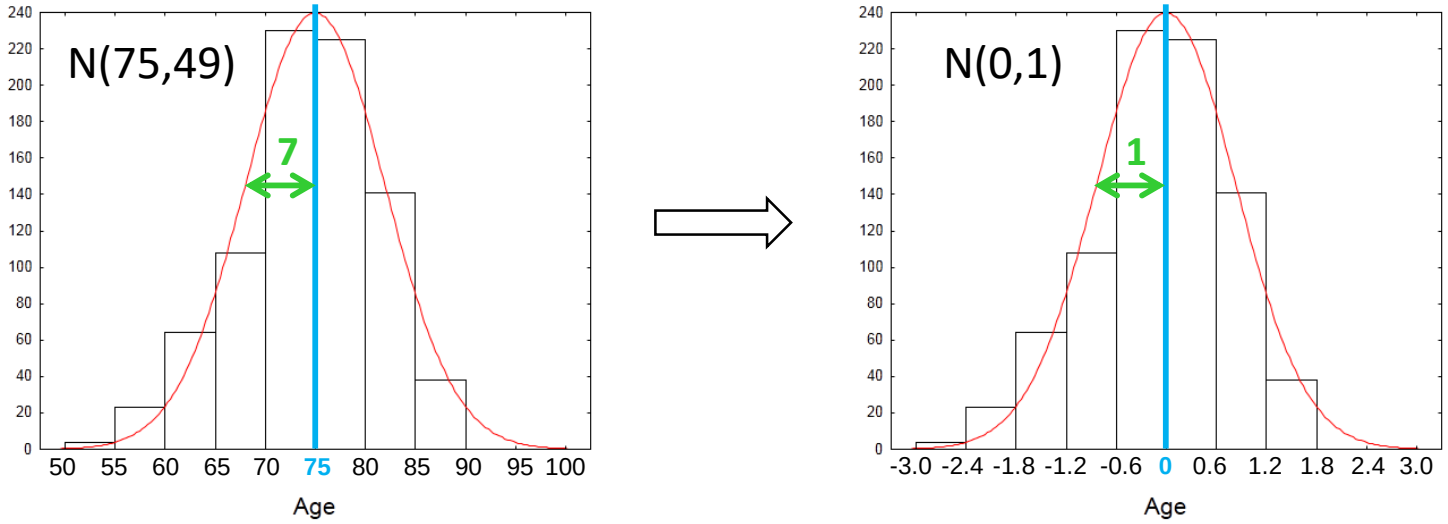
d(l) – šířka intervalu
n(l) – absolutní četnost
n(l) / n – intervalová relativní četnost
N(x'') – intervalová kumulativní četnost do horní hranice x''
F(x'') – intervalová relativní kumulativní četnost do horní hranice X''

Normální rozdělení – různé μ a σ^2



Standardizované normální rozdělení

- Jakékoliv normální rozdělení může být převedeno na tzv. standardizované normální rozdělení: $X \sim N(\mu, \sigma^2) \rightarrow Z = \frac{X - \mu}{\sqrt{\sigma^2}} \sim N(0,1)$

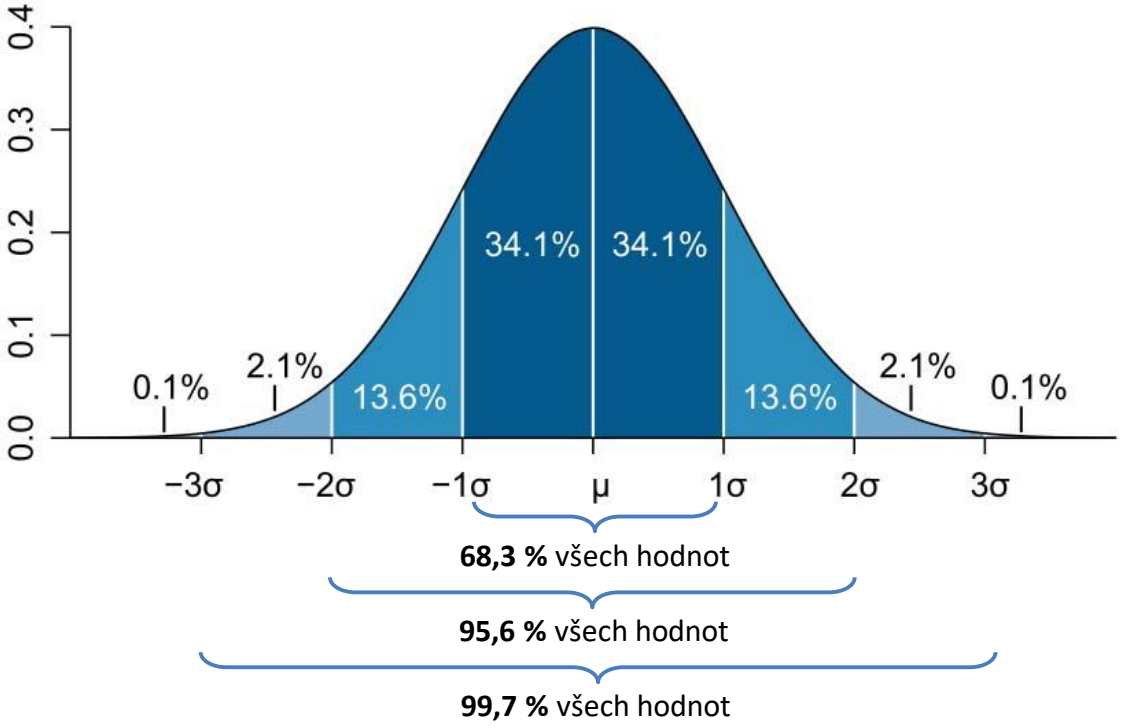


→ **střední hodnota rovna 0, rozptyl roven 1**

- Hustota pravděpodobnosti: $f(z;0,1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$
- **Klíčové rozdělení řady testů.**
- Výhoda je, že všechny hodnoty distribuční i kvantilové funkce jsou tabelovány a obsaženy ve všech dostupných softwarech.

Normální rozdělení – pravidlo ±3 sigma

- U normálního rozdělení lze vyčíslit procento hodnot, které by se měly vyskytovat v rozmezí ± *x* násobku směrodatné odchylky (SD=σ) od průměru.
- Lze říci, že v rozmezí μ ± 3σ by se mělo vyskytovat přes 99,5 % všech hodnot.

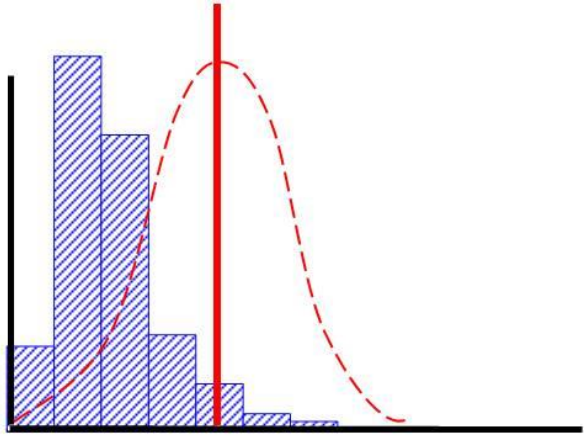


- Použití: orientační ověření normality dat, identifikace odlehlých hodnot

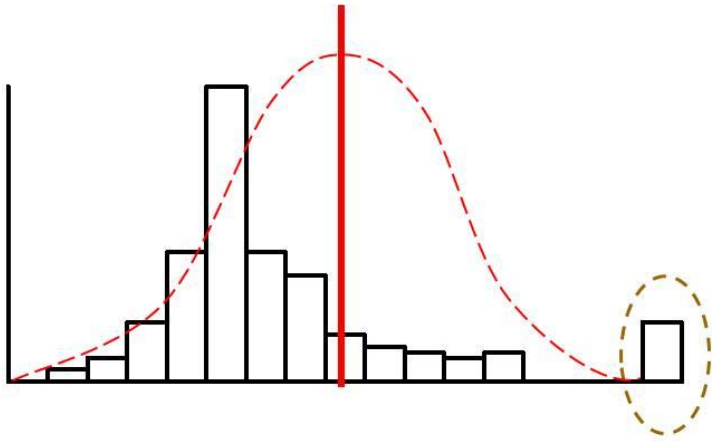
Normalita dat

- Normalita je klíčovým předpokladem řady statistických metod – zejména testů a modelů.
- Není-li splněna podmínka normality hodnot, je špatně celý model, se kterým daná metoda pracuje, což vede k neinterpretovatelným závěrům.
- Její ověření je tak stejně důležité jako výběr správného testu.
- Pro ověření normality existuje řada testů a grafických metod.

Rozdělení není normální

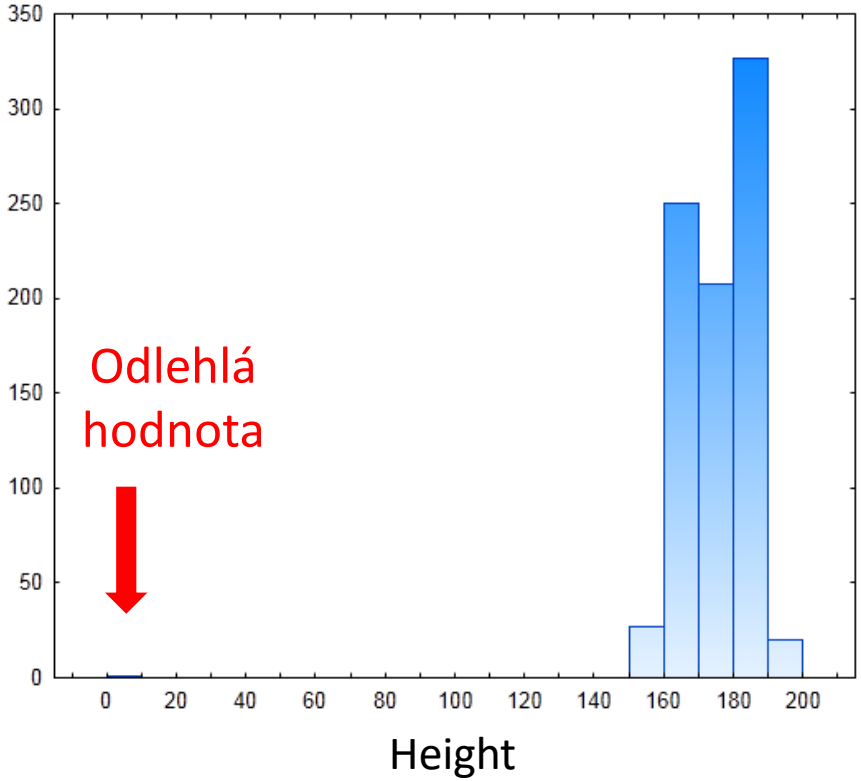


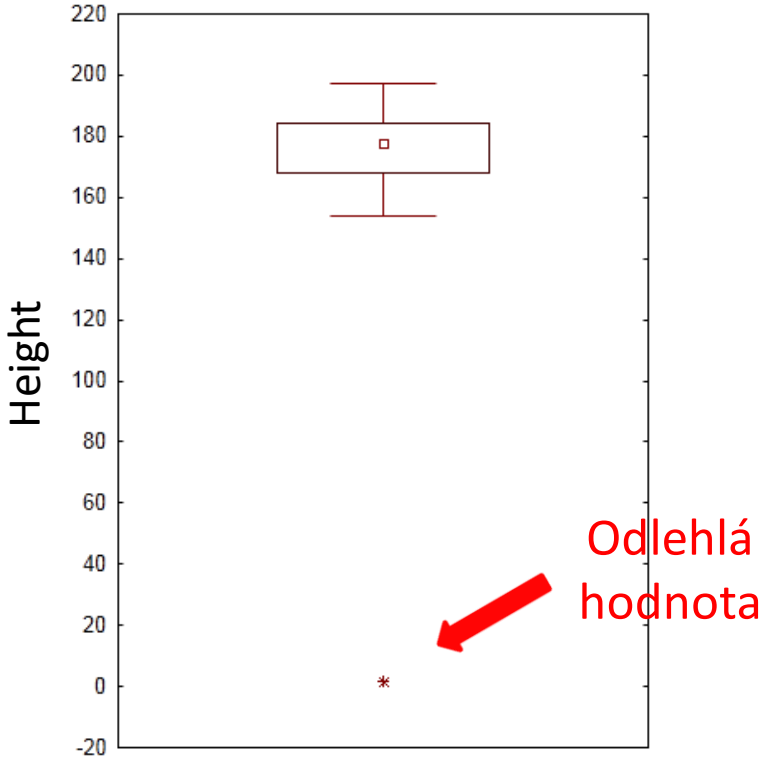
Odlehlá hodnota



Odlehlá hodnota

- Netypické pozorování
- Závisí však na naší znalosti dané problematiky, jestli je daná hodnota možná či nikoliv!
- Grafická identifikace: pomocí histogramu a krabicového grafu





Odlehlá hodnota

- Identifikace pomocí popisných statistik: srovnání mediánu a průměru a pomocí směrodatné odchylky

	Valid N	Mean	Median	Minimum	Maximum	Std.Dev.
Height	833	176.0	178.0	1.6	197.0	11.0
Height_cor	833	176.2	178.0	154.0	197.0	9.2

	Valid N	Mean	Median	Minimum	Maximum	Std.Dev.
Height	20	166.3	174.0	1.6	193.0	39.6
Height_cor	20	174.2	174.0	158.0	193.0	8.9

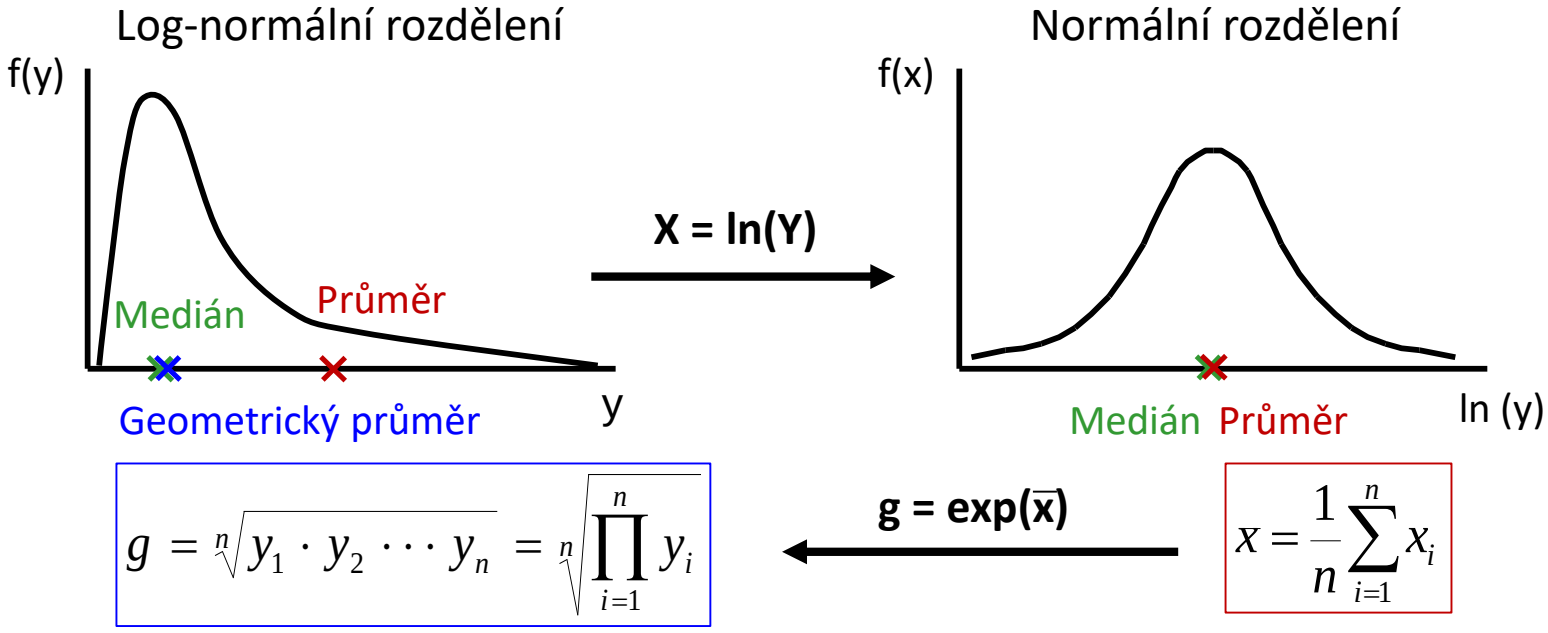
- U velkého datového souboru bude průměr méně ovlivněn odlehlou hodnotou, z popisných statistik nemusíme poznat, že by tam mohla být odlehlá hodnota -> vždy provádět vizualizaci dat!

Úkol 3

- Zjistěte, zda má MMSE skóre normální rozdělení – použijte histogram, krabicový graf a popisnou statistiku.

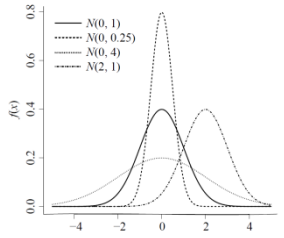
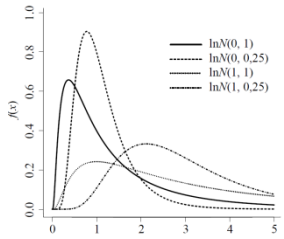
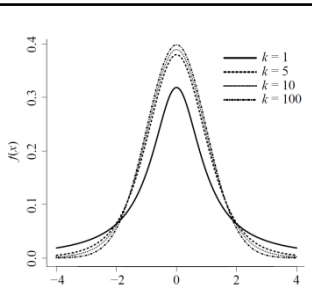
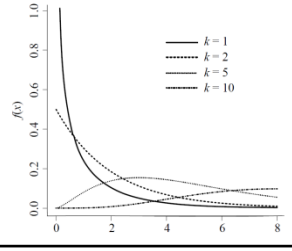
Logaritmicko-normální rozdělení

- u zešikmeného rozdělení nám často (ale ne vždy!) může pomoci proměnnou transformovat pomocí **logaritmické transformace:** X = ln(Y) nebo X = ln(Y+1), pokud data obsahují 0

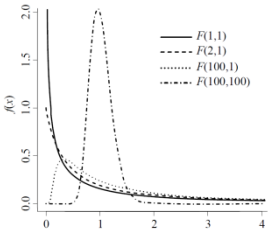
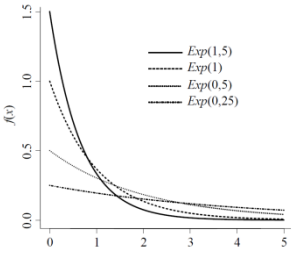
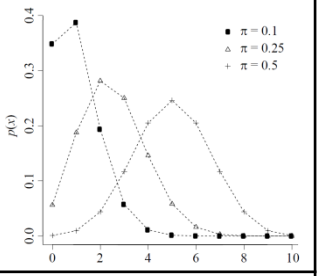
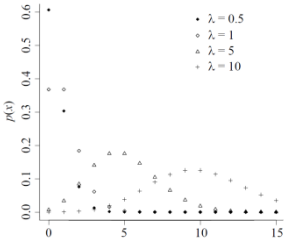


- můžeme použít přirozený logaritmus (ln), dvojkový logaritmus (log2) nebo dekadický logaritmus (log10)
- Příklady veličin s log-normálním rozdělením: tělesná hmotnost, délka inkubační doby infekčního onemocnění, řada krevních parametrů (např. počet krevních buněk v daném objemu krve, sérový bilirubin u pacientů s cirhózou), počet bakteriálních buněk v daném objemu,…

Stručný přehled rozdělení I.

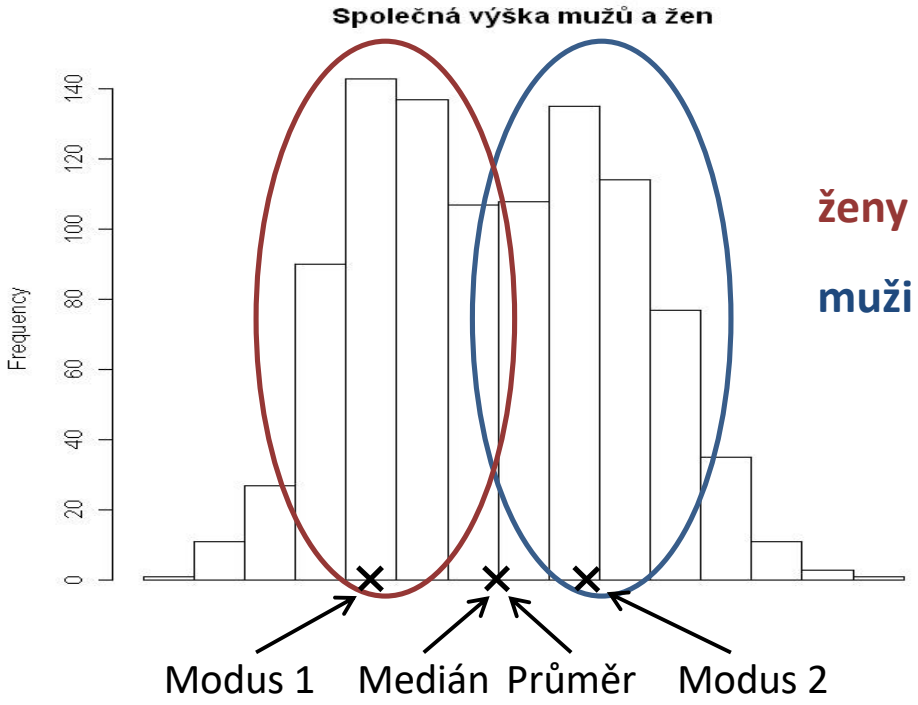
Rozdělení	Parametry	Popis	Graf
Normální $N(\mu,\sigma^2)$	Průměr Rozptyl	Praktická významnost, spojité. EX=μ, DX=σ^2 Př. délkové rozměry těla	
Log-normální $lnN(\mu,\sigma^2)$	Geometrický průměr Rozptyl	Praktická významnost, spojité. EX= $e^{\mu+\sigma^2/2}$, DX= $\left(e^{\sigma^2}-1\right)e^{2\mu+\sigma^2}$ Př. objemové rozměry, hmotnost	
Studentovo t $t(k)$	Stupně volnosti (uvažuje velikost vzorku) Průměr, Rozptyl	Teoretická významnost, spojité. Aproximace normálního rozd. pro malé soubory, pro větší soubory (n>100) se limitně blíží normálnímu rozd. Teoretický základ *t* testu.	
Chí-kvadrát $\chi^2(k)$	Stupně volnosti (uvažuje velikost vzorku)	Teoretická významnost, spojité. Porovnávání četností jevů ve 2 a více kategoriích, výpočet intervalu spolehlivosti pro rozptyl.	

Stručný přehled rozdělení II.

Rozdělení	Parametry	Popis	Graf
Fisherovo F $F(k_1,k_2)$	Dvojí stupně volnosti (uvažuje velikost dvou vzorků)	Teoretická významnost, spojité. Základ ANOVA testu a F-testu, výpočet intervalu spolehlivosti pro podíl rozptylů.	
Exponenciální *Exp*(λ)	Průměr Rozptyl	Praktická význ., spojité. EX= 1/λ, DX=1/λ^2 Popisuje dobu mezi událostmi, význam v analýze přežití, zobecněním je Weibullovo a Gamma rozdělení. Př. doba od diagnózy do úmrtí	
Binomické *Bi*(n,π)	Průměr Rozptyl	Praktická významnost, diskrétní. EX=nπ, DX=nπ(1-π) Popisuje počet výskytů sledované události v n nezávislých pokusech. Př. výskyt nežádoucích účinků léků.	
Poissonovo *Po*(λ)	Průměr Rozptyl	Praktická významnost, diskrétní. EX= λ, DX=λ Popisuje počet výskytů sledované události na danou jednotku času, plochy... Př. počet krvinek v poli mikroskopu.	

Bimodální rozdělení

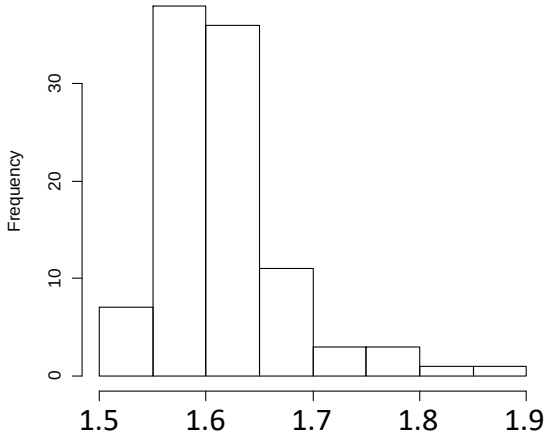
- Představuje většinou problém, neboť se zřejmě jedná o směs dvou souborů s unimodálním rozdělením.
- Bimodální rozdělení má např. tento tvar:



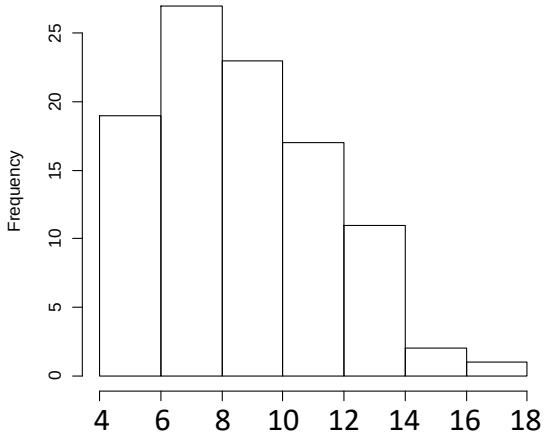
- Nutná další analýza: Co způsobuje bimodalitu? Umožňuje proměnná rozlišit kategorie lidí (např. pacienty od kontrol)? Je vzorek reprezentativní?

Úkol 4 - Přiřaďte k daným veličinám jejich název a typ rozdělení.

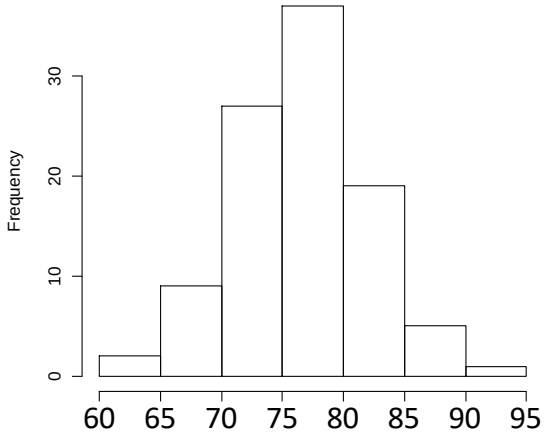
X1: 1.58 1.55 1.67 1.69 1.57



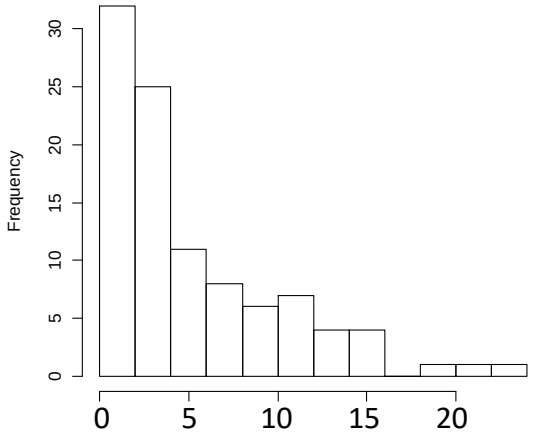
X2: 10 12 8 7 10



X3: 79.5 89.2 75.3 77.8 90.0



X4: 0.49 0.78 6.01 0.47 4.70



Vybraná rozdělení:

I. Normální rozdělení
II. Logaritmicko-normální rozdělení
III. Poissonovo rozdělení
IV. Exponenciální rozdělení

Veličiny:

a) Doba od zahájení léčby do kompletní remise u pacienta s chronickou myeloidní leukémií (v letech)
b) Plocha kůže člověka (v m^2)
c) Diastolický tlak (v mm Hg)
d) Počet příjezdů sanitky do okresní nemocnice za hodinu

4. Transformace dat

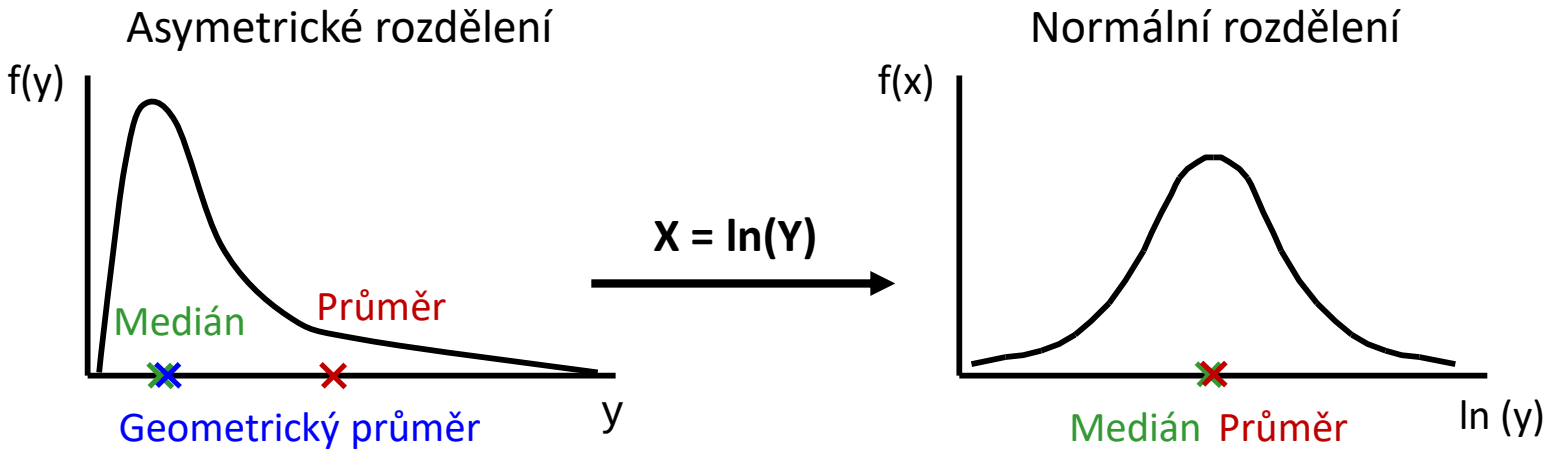
Význam transformací

- Transformace umožní změnit rozsah hodnot proměnné, změnit typ rozložení apod.

- Hlavní cíle transformací:
  1. Normalizace dat – převod na normální rozdělení
  2. Standardizace dat – převod na standardizované normální rozdělení
  3. Centrování dat
  4. Lepší interpretace dat

Normalizace dat

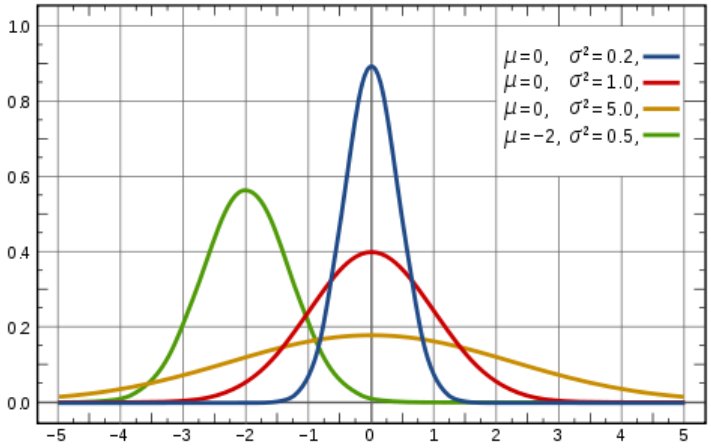
- Převod na normální rozdělení (normalita je předpokladem řady statistických testů).
- Např. **logaritmická transformace**: X = ln(Y) nebo X = ln(Y+1), pokud data obsahují hodnotu 0



- Další příklady:
  - **odmocninová transf.** (pro proměnné s Poissonovým rozložením nebo obecně data typu počet jedinců, buněk apod.: $X = \sqrt{Y}$ nebo $X = \sqrt{Y+1}$
  - **arcsin transfomace** (pro proměnné s binomickým rozložením)
  - **Box-Coxova tranformace**

Standardizace dat

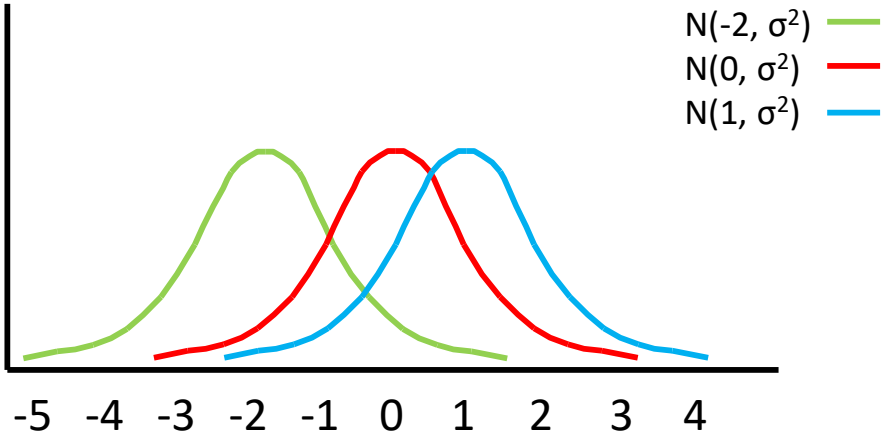
- Převod proměnné s normálním rozdělením na standardizované normální rozdělení: $N(\mu,\sigma^2) \rightarrow N(0,1)$
- Důvod: řada statistických metod byla odvozena pro standardizované normální rozdělení, N(0,1). Děláme to tedy opět kvůli lepší možnosti hodnocení dat.
- Standardizace: $u_i = \frac{x_i - \bar{x}}{s}$
- Obrázek – standardizace je převod „modré“, „zelené“ a „okrové“ na „červenou“.



- z-skóre vlastně vyjadřuje, o kolik směrodatných odchylek se i-tá hodnota odchýlila od průměru.

Centrování dat

- Odečtení průměru od dat – získáme novou proměnnou, která bude mít střední hodnotu rovnu nule: $N(\mu,\sigma^2) \rightarrow N(0, \sigma^2)$
- Důvod: Centrování je důležitou podmínkou některých pokročilých statistických metod (např. klasifikačních).
- Centrování: $u_i = x_i - \bar{x}$
- Obrázek – centrování je převod „modré“ a „zelené“ na „červenou“.



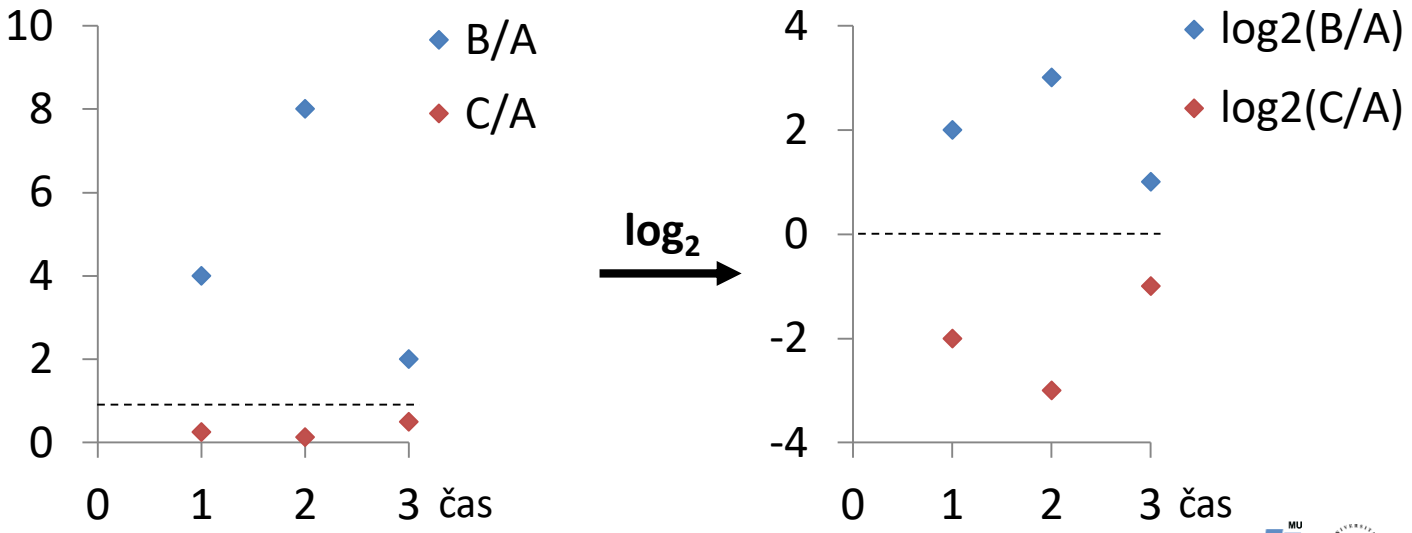
Transformace kvůli lepší interpretaci dat

- Příklad: Microarray experiment se dvěma vzorky, měříme intenzitu exprese genu XY v jedné tkáni (hodnota intenzity A_{XY}) a v druhé tkáni (hodnota intenzity B_{XY}).
- Následně hodnoty převádíme na logaritmus se základem 2 jejich podílu:

$$Z_{XY} = \log_2\left(\frac{B_{XY}}{A_{XY}}\right)$$

- Umožní nám to posoudit kolikrát byla exprese jednoho genu větší/menší než druhého genu (2x, 4x, 8x, 16x,....).

čas	B/A	C/A
1	4	1/4
2	8	1/8
3	2	1/2



Další příklady transformací – odvozené typy dat

- **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
- **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.
- **Skóre** (score) – jedná se o uměle vytvořené hodnoty charakterizující určitý stav, který nelze jednoduše měřit jako číselné hodnoty. Př.: indexy kvality života.

Kategorizace

- Vytvoření kvalitativní proměnné z kvantitativní proměnné.

Primární data

Age
84
76
79
89
71
70
88
86
.
.
.
.
.
.
n=833

Kategorizace →

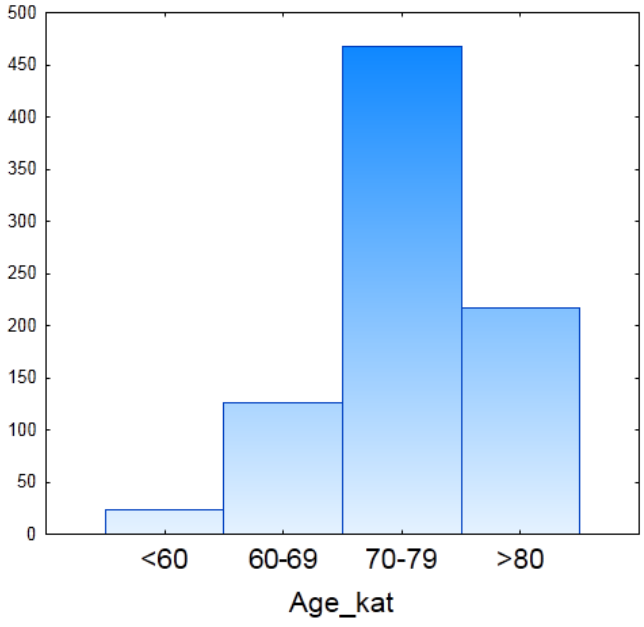
Frekvenční tabulka

	n(x)	N(x)	p(x)	F(x)
<60	23	23	2,8	2,8
60-69	126	149	15,1	17,9
70-79	467	616	56,1	73,9
>80	217	833	26,1	100,0

x: Kategorizovaný věk

n(x) – absolutní četnost x
N(x) – kumulativní četnost hodnot nepřevyšujících x; $N(x) = \sum_{t \le x} n(t)$
p(x) – relativní četnost; $p(x) = n(x) / n$
F(x) – kumulativní relativní četnost hodnot nepřevyšujících x; $F(x) = N(x) / n$

Vizualizace



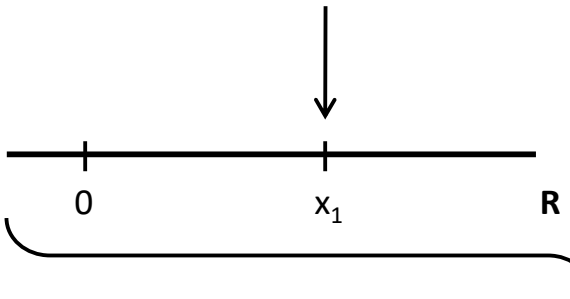
Úkol 5

- Vytvořte novou proměnnou, která bude obsahovat standardizovaný objem amygdaly.
- Vytvořte novou proměnnou, která bude obsahovat kategorizovanou váhu (kategorie zvolte na základě histogramu).

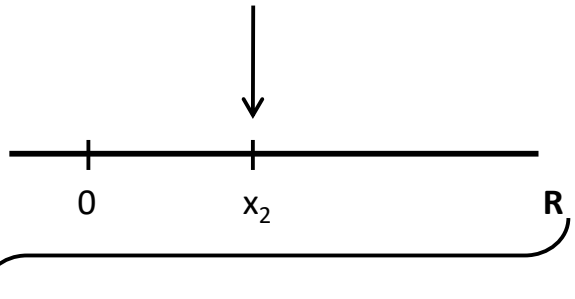
5. Intervaly spolehlivosti

Intervaly spolehlivosti – motivace

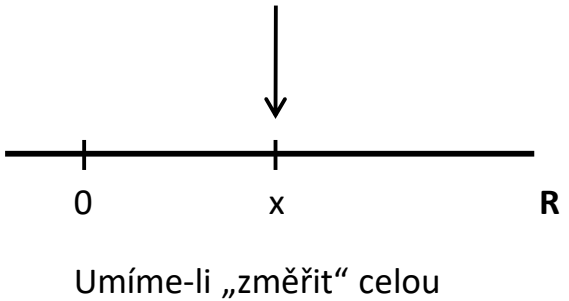
Výběr číslo 1



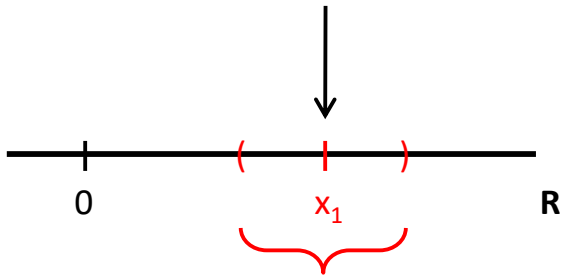
Výběr číslo 2



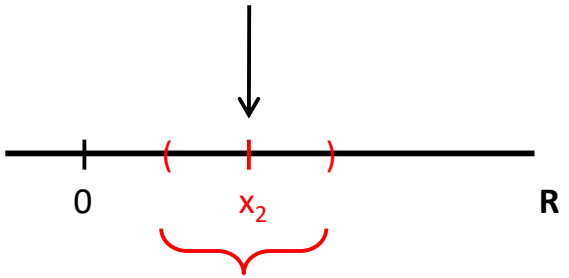
Celá cílová populace



Pracujeme-li s výběrem z cílové populace, je třeba na základě variability pozorovaných dat spočítat tzv. interval spolehlivosti pro bodový odhad.



Interval spolehlivosti na základě výběru číslo 1.

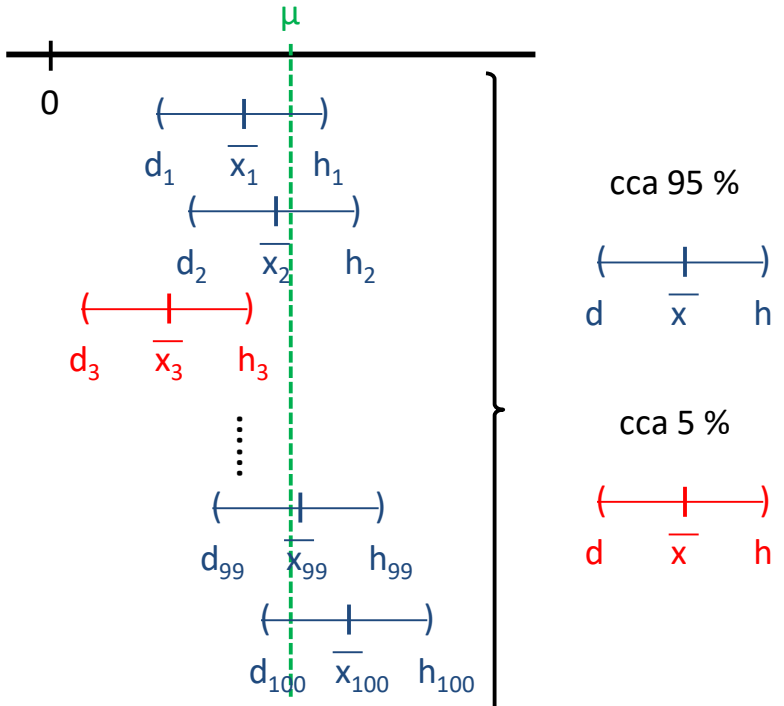


Interval spolehlivosti na základě výběru číslo 2.

Umíme-li „změřit“ celou cílovou populaci, nepotřebujeme interval spolehlivosti, protože jsme schopni odhadnout sledovaný parametr přesně – v praxi je tato situace nereálná.

Interval spolehlivosti (IS) – interpretace

- Interval spolehlivosti ukazuje, jak přesný je výpočet průměru.
- 95% interval spolehlivosti vymezuje prostor kam s 95% pravděpodobností padne populační průměr vypočtený při dalším vzorkování populace (za stejných podmínek a o stejné velikosti vzorku). **Tedy 95% interval spolehlivosti obsahuje populační průměr s rizikem α=0,05 (5%)**.
- Čím je interval spolehlivosti užší, tím přesnější je náš odhad průměru (tím víc se náš odhad průměru pomocí našeho vzorku blíží populačnímu průměru).
- 95% interval spolehlivosti - ilustrace: Pokud bychom opakovaně vybírali skupiny subjektů o stejné velikosti a počítali průměr a interval spolehlivosti, tak 95% intervalů spolehlivosti by pokrývalo populační průměr μ a 5% intervalů spolehlivosti by populační průměr nepokrývalo.



Střední chyba průměru

- Nebo též standardní chyba průměru („standard error“) – značka *SE*.
- **Neplést se *SD* (směrodatnou odchylkou)!!!**
- $SE = \frac{SD}{\sqrt{n}}$
- *SE* je založena na směrodatné odchylce dat a počtu hodnot (vlastně jde o směrodatnou odchylku rozložení průměru).
- Říká, jak přesný je výpočet průměru:
  - velký počet subjektů (*n*), z nichž počítáme průměr → tím menší je *SE* (tzn. tím přesnější je průměr)
  - malý počet subjektů (*n*), z nichž počítáme průměr → tím větší je *SE* (tzn. tím méně přesný je průměr)

Interval spolehlivosti - poznámka

- Interval spolehlivosti (Confidence Interval – CI)
- Interval spolehlivosti pro průměr se tedy vypočítá jako:

$$x - SE \cdot 1{,}96 \leq \mu \leq x + SE \cdot 1{,}96$$

- Interval spolehlivosti má smysl počítat pouze v případě, že mají data normální rozdělení!

- Interval spolehlivosti počítá pouze s variabilitou danou náhodným výběrem, nepočítá se zdroji systematického zkreslení – např.
  - Měření krevního tlaku může být systematicky zkresleno starým měřidlem („technical bias“).
  - Měření krevního tlaku může být systematicky zkresleno tím, že se do studie přihlásí pouze určitá skupina osob („selection bias“).

Interval spolehlivosti pro μ

$$P(D \leq odhad \leq H) > 1 - \alpha$$

Obecný tvar intervalu spolehlivosti (IS):

Odhadovaný parametr	±	Chyba odhadu	*	Kvantil modelového rozložení pro (1-a/2)

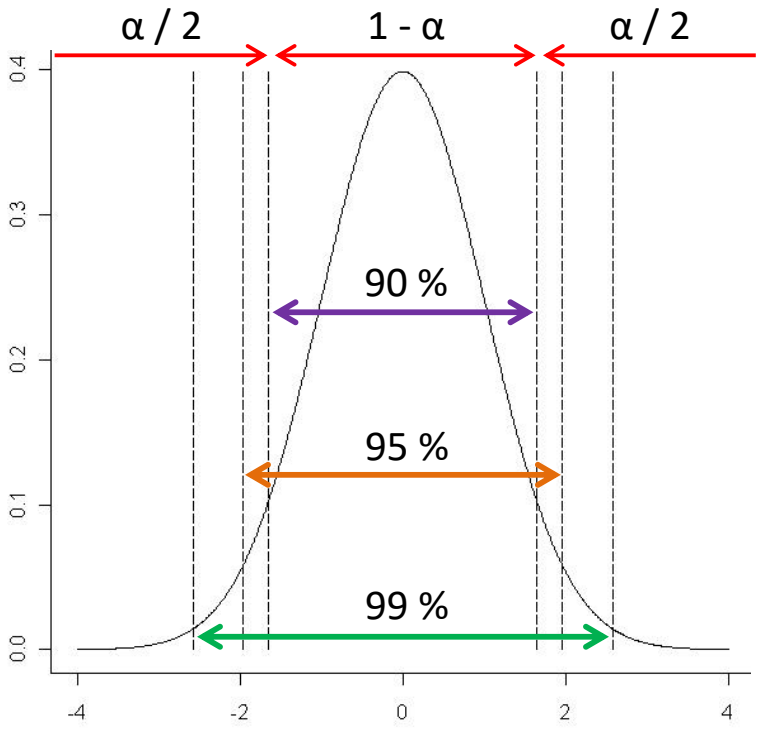
Interval spolehlivosti pro μ:

$$\bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \leq \mu \leq \bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}$$

dolní mez IS (D) — horní mez IS (H)

- $\bar{x}$ ... výběrový průměr
- σ ... směrodatná odchylka
- n ... velikost výběrového souboru
- $z_{1-\alpha/2}$ ... kvantil standardizovaného normálního rozdělení
- α ... riziko
- $\frac{\sigma}{\sqrt{n}}$ ... **střední chyba odhadu průměru**

Kvantily standardizovaného normálního rozdělení



$z_{0,005} = -2,58$ $\quad 2,58 = z_{0,995}$

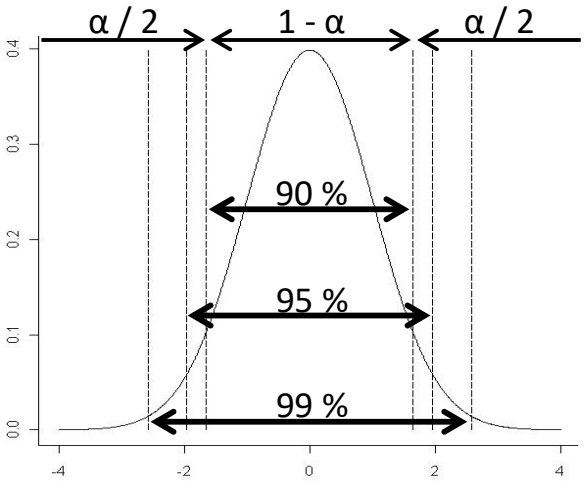
$z_{0,025} = -1,96$ $\quad 1,96 = z_{0,975}$

$z_{0,050} = -1,64$ $\quad 1,64 = z_{0,950}$

Ovlivnění šířky intervalu spolehlivosti

- Interval spolehlivosti: $\bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \leq \mu \leq \bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}$

- Co ovlivňuje šířku intervalu spolehlivosti?
  - **Velikost vzorku** – s rostoucí velikostí vzorku je IS užší (máme více informace, a tak je odhad přesnější)
  - **Variabilita náhodné veličiny** – čím náhodná veličina vykazuje větší variabilitu, tím je IS pro odhad střední hodnoty širší, tedy odhad je méně přesný.
  - **Spolehlivost, kterou požadujeme** – s rostoucí spolehlivostí (tzn. menším α), je IS širší, neboť požadujeme větší jistotu, že náš interval skutečně pokrývá hodnotu neznámého parametru). Standardně se používá 95% IS (odpovídající riziku α=5%), ale v literatuře se lze setkat i s 90% anebo 99% IS (99% IS tedy bude širší než 95% IS).



Interval spolehlivosti pro μ při neznámém σ

- IS pro μ při známém σ: $\bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \leq \mu \leq \bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}$
- **IS pro μ při neznámém σ**: $\bar{x} - \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1) \leq \mu \leq \bar{x} + \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1)$
- Přesnou hodnotu populační σ v praxi většinou neznáme → snažíme se ji odhadnout pomocí výběrové směrodatné odchylky s: $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- $t_{1-\propto/2}(n-1)$ je kvantil Studentova t rozdělení

- **Příklad:** V našem souboru má 833 lidí průměrný věk roven 74,8 let a směrodatná odchylka věku je 6,9 let. Vypočtete 95% IS pro odhad střední hodnoty věku.
- Řešení:

$n = 833$
$\bar{x} = 74{,}8$ let
$s = 6{,}9$ let

$$\bar{x} - \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1) \leq \mu \leq \bar{x} + \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1)$$

$$74{,}8 - \frac{6{,}9}{\sqrt{833}} t_{1-0{,}05/2}(833-1) \leq \mu \leq 74{,}8 + \frac{6{,}9}{\sqrt{833}} t_{1-0{,}05/2}(833-1)$$

$$74{,}3 \leq \mu \leq 75{,}3$$

Další druhy intervalů spolehlivosti

- Interval spolehlivosti **pro rozdíl průměrů dvou výběrů** (jde nám např. o srovnání objemu hippocampu u pacientů a kontrol):

$$\overline{X} - \overline{Y} - t_{1-\alpha/2}(n_1 + n_2 - 2)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \overline{X} - \overline{Y} + t_{1-\alpha/2}(n_1 + n_2 - 2)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- Interval spolehlivosti **pro odhad rozptylu**:

$$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}$$

- Interval spolehlivosti **pro podíl rozptylů dvou výběrů** (lze ho použít pro hodnocení homogenity rozptylů dvou výběrů, která je jedním z předpokladů v testování hypotéz):

$$\frac{s_2^2}{s_1^2} F_{\alpha/2}(n_1 - 1, n_2 - 1) \leq \frac{\sigma_2^2}{\sigma_1^2} \leq \frac{s_2^2}{s_1^2} F_{1-\alpha/2}(n_1 - 1, n_2 - 1)$$

- Druhů intervalů spolehlivosti je ještě mnohem více – např. IS pro medián, pro podíl,...

Neparametrické metody pro konstrukci IS

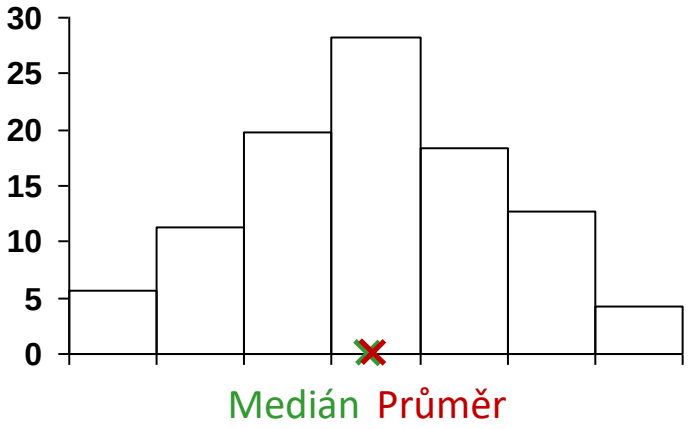
- **Bootstrap** – je založen na principu opakovaného vzorkování naměřených dat s vracením, kdy pro vytvoření nového vzorku dat může být každý prvek použit více než jednou, právě jednou anebo není použit vůbec (ovšem se zachováním celkové velikosti souboru *n* i velikosti jednotlivých skupin). Pro každý vzorek je vypočítán výběrový průměr, tyto výběrové průměry seřadíme podle velikosti a vypočítáme 2,5% a 97,5% kvantil (stejně jako jsme počítali 80% kvantil na slidu 32), které nám dají dolní a horní mez pro 95% IS.
- **Jackknife** – opakovaný výpočet sledované charakteristiky je prováděn vždy s vynecháním právě jednoho pozorování. Tento postup nám stejně jako v případě metody bootstrap poskytuje představu o rozsahu hodnot, ve kterých se námi sledovaná charakteristika může pohybovat, budeme-li považovat naměřená data za reprezentativní vzorek z cílové populace.

Úkol 6

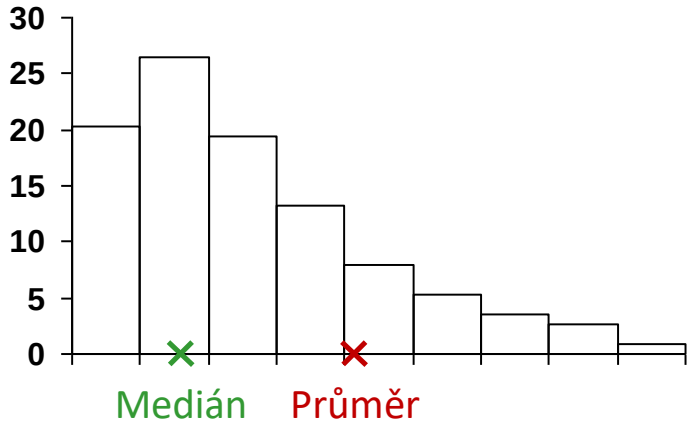
- Vypočtěte průměr, střední chybu průměru a intervaly spolehlivosti pro všech šest mozkových struktur a MMSE skóre.
- Zamyslete se nad tím, zda mělo vůbec smysl počítat intervaly spolehlivosti pro všechny výše uvedené proměnné.

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
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“