# Vícerozměrné statistické metody

Smysl a cíle vícerozměrné analýzy dat a modelování, vztah jednorozměrných a vícerozměrných statistických metod

Jiří Jarkovský, Simona Littnerová, Eva Koriťáková

# Průběh výuky

- přednášky doplněné o praktické cvičení v SW

- Obsahem předmětu je přehled a úvod do praktické aplikace vícerozměrných statistických metod:
  - Shluková analýza
  - Ordinační metody
  - Základy vícerozměrné diskriminace

- Předpoklady ukončení
  - Účast na cvičení (jedna absence povolena)
  - průběžné odpovědníky v IS.MUNI
  - písemná závěrečná zkouška (požadováno dosažení alespoň 50% bodů)

# Plán přednášek

18.9. Odpadá

25.9. Smysl a cíle vícerozměrné analýzy dat a modelování, vztah jednorozměrných a vícerozměrných statistických metod

2.10. Vícerozměrné statistické rozdělení a testy, operace s vektory a maticemi + Podobnosti a vzdálenosti ve vícerozměrném prostoru, asociační matice I

9.10. Cvičení – opakování jednorozměrné analýzy dat

16.10. Cvičení – asociační matice + shluková analýza

23.10. Podobnosti a vzdálenosti ve vícerozměrném prostoru, asociační matice II + Shluková analýza – hierarchická, nehierarchická + Identifikace optimálního počtu shluků

30.10. Ordinační analýzy – principy redukce dimenzionality - PCA+FA, Ordinační analýzy – CA, DCA + MDS

6.11. Ordinační analýza – vztah ordinačních prostorů (RDA, CCA, co-inertia, pouze přehled) + Diskriminační analýza

13.11. Cvičení - PCA

20.11. Cvičení – CA, DCA, CCA + diskriminační analýza

27.11. Rezerva

4.12. Opakovací přednáška

11.12. Předtermín

# Vícerozměrné statistické metody

Smysl a cíle vícerozměrné analýzy dat

# Význam a cíle vícerozměrné analýzy dat

- většina dat pořízených při výzkumu jsou data vícerozměrná – chceme zjistit celou řadu vlastností daných subjektů či objektů

PROMĚNNÉ (VLASTNOSTI)

| SUBJEKTY | ID | Pohlaví | Věk | Váha | MMSE skóre | Objem hipokampu | … |
|---|---|---|---|---|---|---|---|
| | 1 | muž | 84 | 85,5 | 29 | 7030 | |
| | 2 | žena | 25 | 62,0 | 28 | 6984 | |
| | … | | | | | | |

- zpravidla nestačí analyzovat každou proměnnou zvlášť – pro úplné pochopení vztahů většinou potřeba analyzovat proměnné současně

→ použití VÍCEROZMĚRNÝCH METOD

# Význam a cíle vícerozměrné analýzy dat II

- vícerozměrné metody umožňují:
  - znázornit a popsat vícerozměrná data
  - zjišťovat vztahy mezi jednotlivými proměnnými a mezi subjekty (resp. objekty)

- mnoho způsobů dělení vícerozměrných metod do skupin – např. dělení podle cíle, kterého chceme vícerozměrnou analýzou dosáhnout:

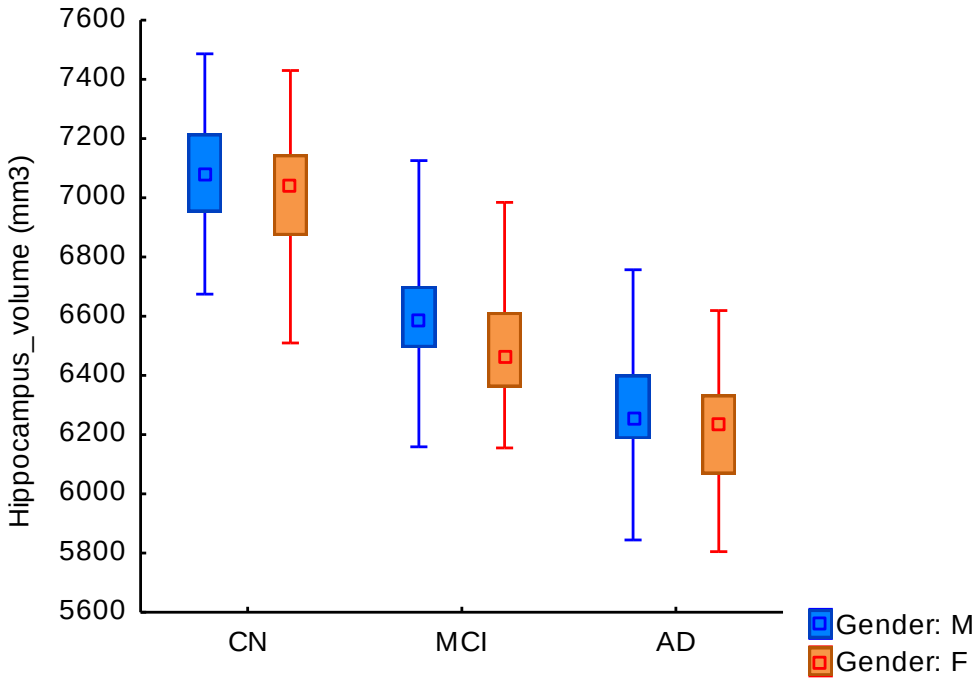
1. Testování hypotéz o vícerozměrných datech
2. Vytvoření shluků subjektů, objektů nebo proměnných
3. Redukce vícerozměrných dat
4. Klasifikace subjektů či objektů

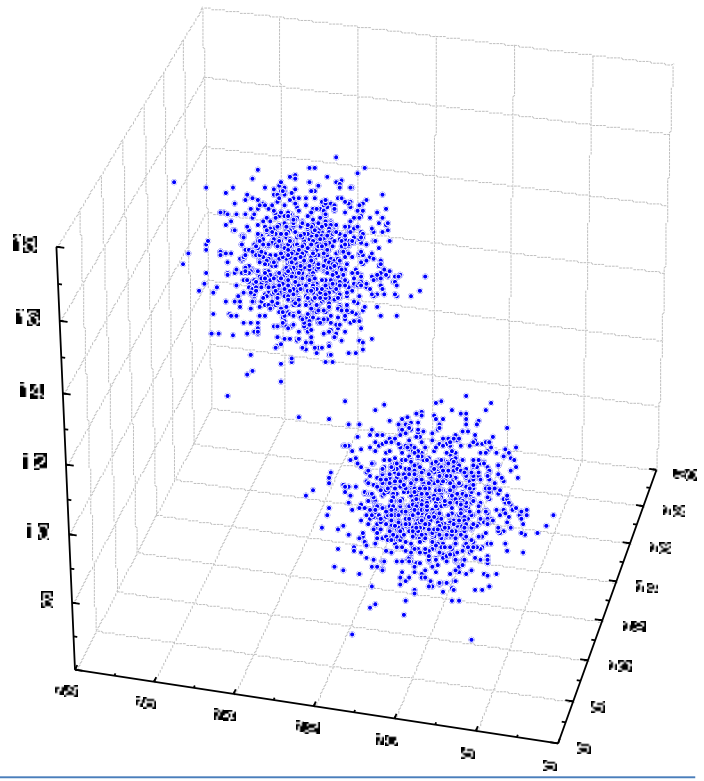
# Cíle vícerozměrné analýzy dat

## 1. Testování hypotéz o vícerozměrných datech

Příklady:

- ověření, zda má vliv pohlaví a typ léku na počet uzdravených pacientů s daným onemocněním
- výzkum vztahu typu onemocnění na objem hipokampu, amygdaly a mozkových komor
- zjištění, zda je rozdílná spotřeba elektrické energie ve městech a na vesnicích během týdne a o víkendu



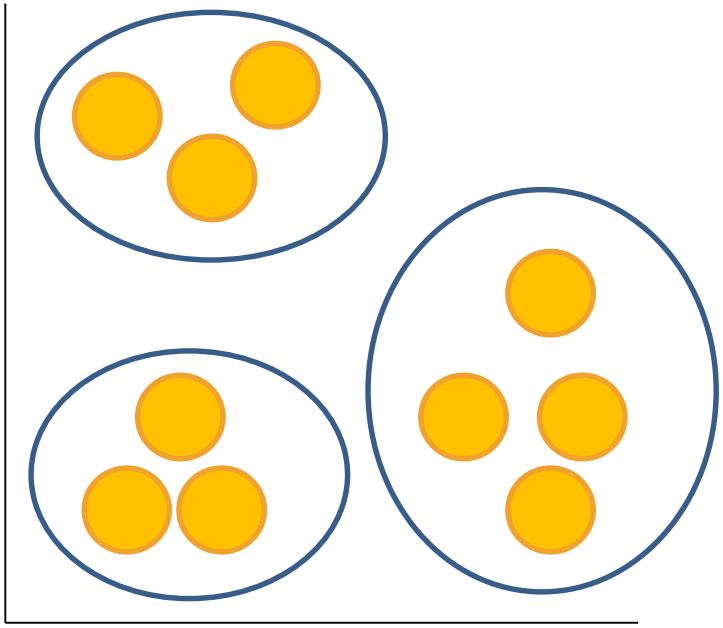



# Cíle vícerozměrné analýzy dat

## 2. Vytvoření shluků subjektů, objektů nebo proměnných

Příklady:

- vytvoření skupin diagnóz onemocnění s podobnými léčebnými náklady
- vytvoření skupin lokalit podle výskytu určitých druhů rostlin a živočichů
- vytvoření skupin genů a subjektů na základě dat genové exprese
- vytvoření skupin subjektů se schizofrenií podle kognitivních skóre a neurologických parametrů



# Cíle vícerozměrné analýzy dat

## 3. Redukce vícerozměrných dat

Příklady:

- vytvoření souhrnného skóre odpovědi pacientů na radioterapii z původních několika proměnných
- vytvoření menšího počtu nových proměnných z původních dat, které nám umožní znázornit vícerozměrná data ve 2-D či 3-D grafech
- výběr oblastí mozku, které nejvíce odlišují pacienty s neuropsychiatrickým onemocněním od zdravých subjektů

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ID | Group | Gender | Age | MMSE | Hippocampus_ | Amygdala_ | Thalamus_ | Pallidum_ | Putamen_ | Nucl_caud_ |
| 2 | 101 | 1 | M | 84 | 28 | 6996 | 2725 | 12800 | 3914 | 11227 | 3528 |
| 3 | 102 | 1 | F | 76 | 29 | 7187 | 2916 | 12277 | 3606 | 11236 | 3773 |
| 4 | 103 | 1 | M | 79 | 30 | 7030 | 2835 | 12906 | 3638 | 11430 | 4294 |
| 5 | 104 | 1 | F | 89 | 30 | 7263 | 2919 | 12432 | 3678 | 11018 | 3585 |
| 6 | 105 | 1 | F | 71 | 30 | 6867 | 2887 | 12383 | 3689 | 11304 | 3723 |
| 7 | 106 | 1 | F | 70 | 30 | 7331 | 3081 | 12415 | 3553 | 11372 | 3969 |
| 8 | 107 | 1 | F | 88 | 30 | 6705 | 2823 | 12575 | 4150 | 11303 | 2886 |
| 9 | 108 | 1 | F | 86 | 28 | 6586 | 2860 | 12454 | 3945 | 11328 | 3741 |
| 10 | 109 | 1 | F | 84 | 29 | 7036 | 3017 | 12361 | 3827 | 11382 | 3737 |

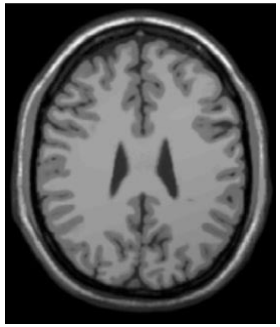
# Cíle vícerozměrné analýzy dat

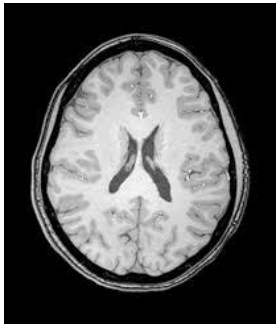
## 4. Klasifikace subjektů či objektů

Příklady:

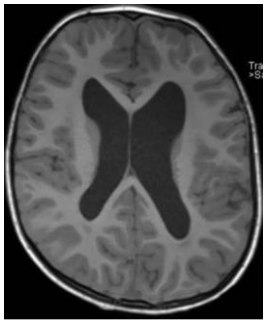
- zjištění (diagnostika) schizofrenie na základě kognitivních testů
- rozhodnutí, zda banka poskytne či neposkytne hypotéku danému subjektu na základě jeho příjmů, rodinné situace atd.
- diagnostika demence (tzn. zařazení nového subjektu do skupiny pacientů či kontrol) podle obrázku mozku

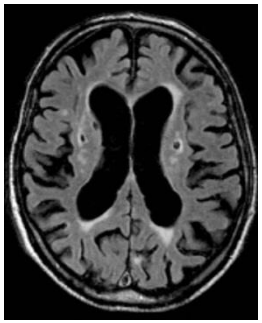
Zdravé subjekty



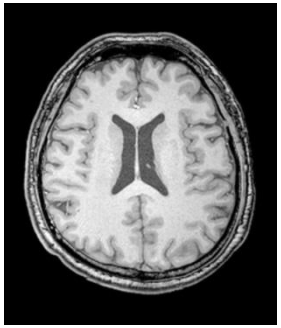


Pacienti





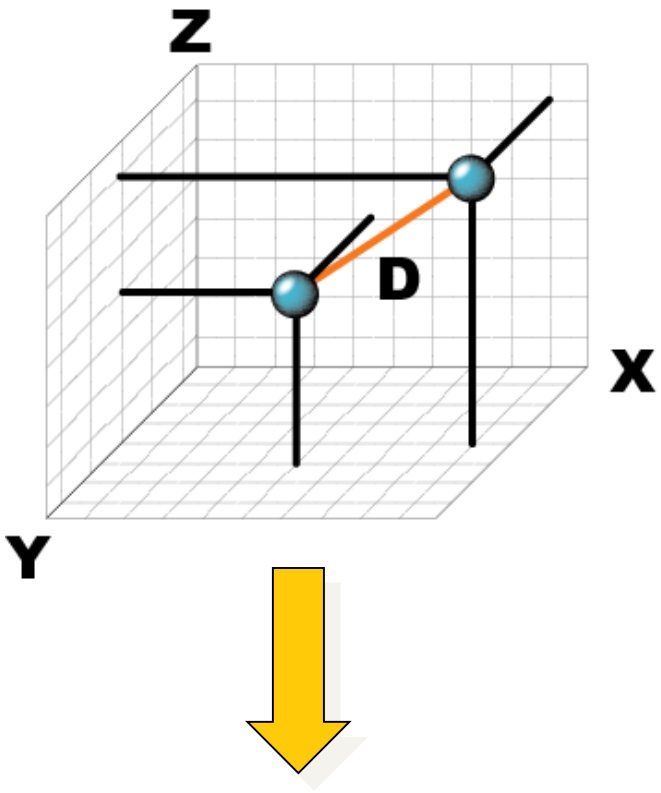
Nový subjekt



Pacient? x Zdravý?

# Cíle vícerozměrné analýzy dat - doplnění

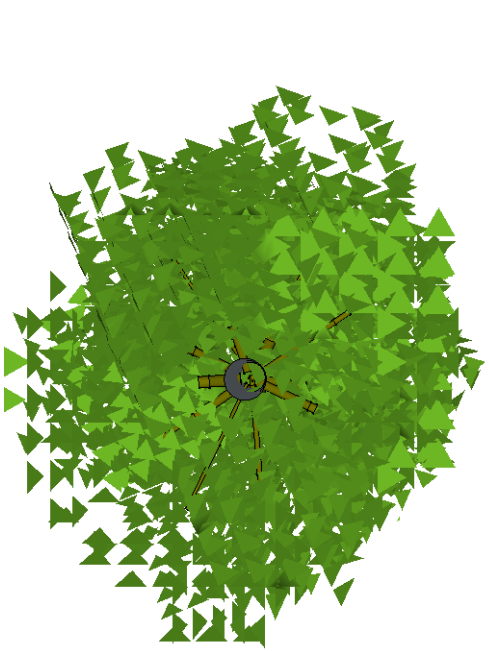
- Každý objekt reálného světa můžeme popsat jeho pozicí v mnohorozměrném prostoru, v extrémním případě jde až o desetitisíce, statisíce či miliony dimenzí
- Více než 3D prostor je pro nás vizuálně neuchopitelný a hledání vztahů ve více než 3 dimenzích je problematické
- Vícerozměrná analýza se tento problém snaží řešit různými přístupy:
  - Redukce dimenzionality dat „sloučením“ korelovaných proměnných do menšího počtu „faktorových“ proměnných
  - Identifikace shluků objektů ve vícerozměrném prostoru a následná redukce vícedimenzionálního problému kategorizací objektů do zjištěných shluků

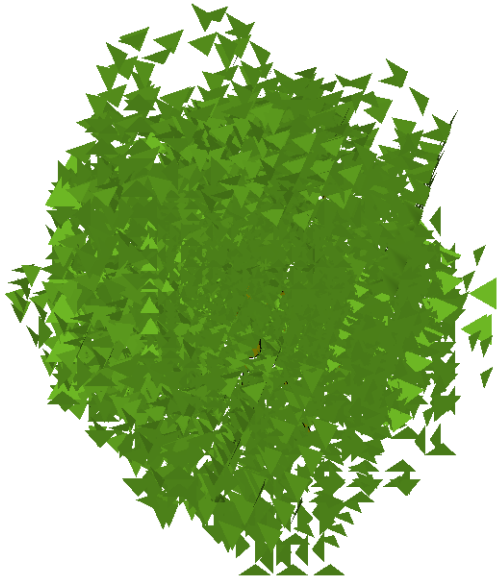


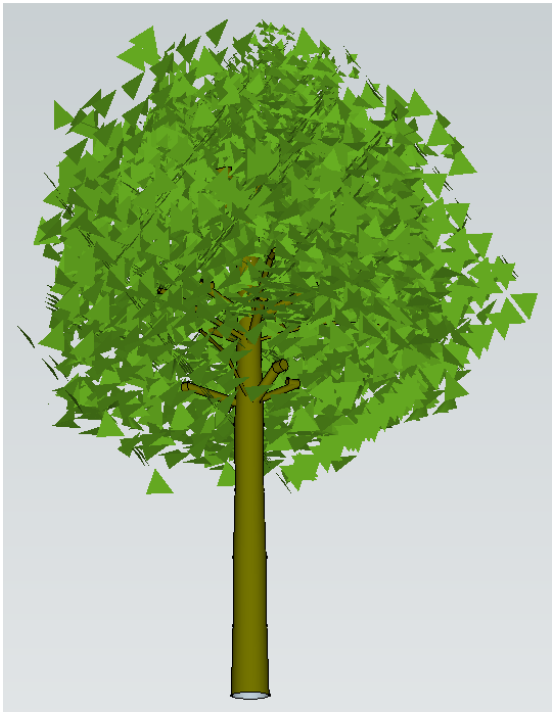

Zjednodušení
Interpretace

# Vícerozměrná analýza dat = pohled ze správného úhlu

- Vícerozměrná analýza nám pomáhá nalézt v x-dimenzionálním prostoru nejvhodnější pohled na data poskytující maximum informací o analyzovaných objektech



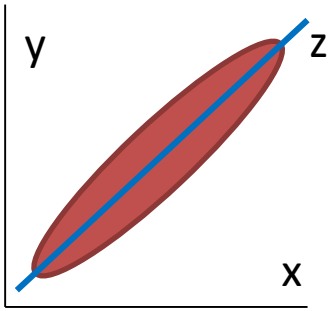




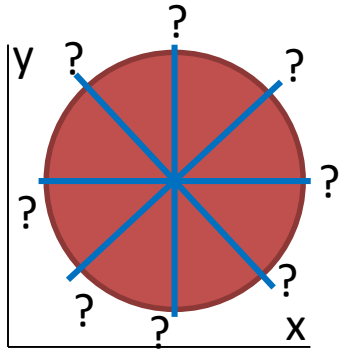
Všechny obrázky ukazují stejný objekt z různých úhlů v 3D prostoru.

# Obecný princip redukce dimenzionality dat

- V převážné většině případů existují mezi dimenzemi korelační vztahy, tedy dimenze se navzájem vysvětlují a pro popis kompletní informace v datech není třeba všech dimenzí vstupního souboru
- Všechny tzv. ordinační metody využívají principu identifikace korelovaných dimenzí a jejich sloučení do souhrnných nových dimenzí zastupujících několik dimenzí vstupního souboru
- Pokud mezi dimenzemi vstupního souboru neexistují korelace, nemá smysl hledat zjednodušení vícerozměrné struktury takovéhoto souboru !!!



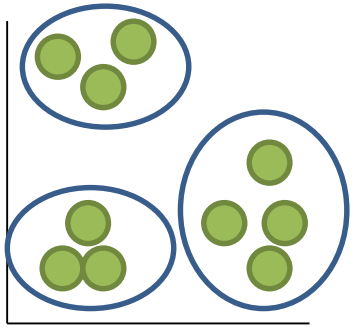

Jednoznačný vztah dimenzí x a y umožňuje jejich nahrazení jedinou novou dimenzí z



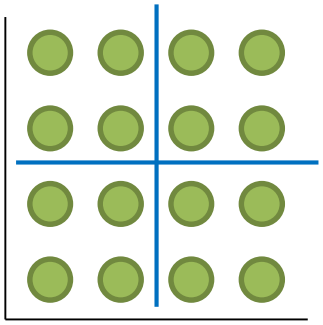

V případě neexistence vztahu mezi x a y nemá smysl definovat nové dimenze – nepřináší žádnou novou informaci oproti x a y

# Obecný princip hledání shluků v datech

- Vzájemnou pozici objektů ve vícerozměrném prostoru lze popsat jejich vzdáleností
- Dle vzdálenosti objektů je můžeme slučovat do shluků a přiřazení objektů ke shlukům ve vícerozměrném prostoru následně využít pro zjednodušení jejich x-dimenzionálního popisu
- Smysluplnost výsledků shlukování závisí jednak na objektivní existenci shluků v datech, jednak na arbitrárně nastavených kritériích definice shluků



Jednoznačné odlišení existujících shluků v datech (obdoba multimodálního rozložení)



Shluková analýza je možná i v tomto případě, nicméně hranice shluků jsou dány pouze naším rozhodnutím.

# Omezení vícerozměrné analýzy dat I

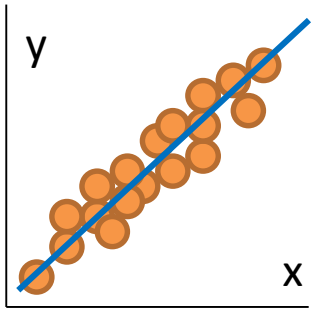
- **Vícerozměrná analýza může přinést zjednodušení dimenzionality dat pouze v případě, kdy data skrývají nějakou identifikovatelnou vícerozměrnou strukturu:**
  - Mezi dimenzemi existují vztahy (korelace) umožňující nahrazení korelovaných dimenzí zástupnou souhrnnou dimenzí
  - Objekty vytváří v x-dimenzionálním prostoru shluky nebo jiné nenáhodné struktury
- Pro náhodně rozmístěné objekty bez korelací mezi dimenzemi jejich x-dimenzionálního prostoru nepřináší vícerozměrná analýza žádné nové informace oproti původním dimenzím!
- **Důležitý je poměr počtu objektů (řádky tabulky) a dimenzí (sloupce tabulky).** Čím je tento poměr menší, tím větší je šance, že výsledky analýzy jsou ovlivněny náhodnými procesy. Za minimální poměr pro získání validních výsledků je považováno 10 objektů na 1 dimenzi.
- **Pro vícerozměrné analýzy platí obdobné předpoklady jako pro jednorozměrnou statistickou analýzu**; vzhledem k jejich možnému porušení na úrovni kombinace několika dimenzí je tyto předpoklady třeba kontrolovat ještě pečlivěji než u jednorozměrné analýzy!

# Omezení vícerozměrné analýzy dat II

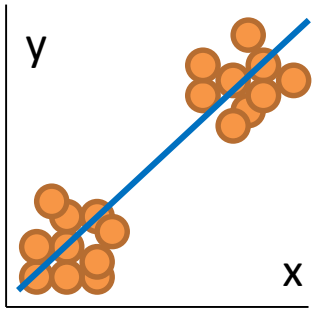
- Kromě klasických statistických předpokladů je při vícerozměrných analýzách třeba věnovat pozornost výběru metrik vzdáleností mezi objekty (klíčové ovlivnění interpretace výsledků) a jejich předpokladům.
- Pokud výsledky vícerozměrné analýzy nejsou interpretovatelné, je třeba zvážit, zda použití vícerozměrné analýzy přináší oproti sadě jednorozměrných analýz nějakou přidanou hodnotou.
- Využitelná vícerozměrná analýza by měla být:
  - Počítána vhodně vybranou metodou pro řešení daného problému
  - Korektně spočítána za dodržení všech předpokladů
  - Interpretovatelná a přinášející novou informaci oproti analýze původních dimenzí

# Korelace jako princip výpočtu vícerozměrných analýz

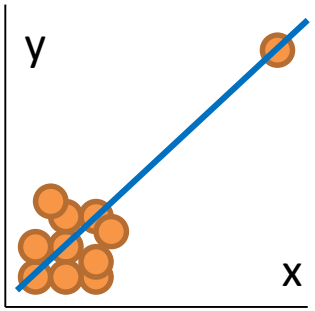
- Kovariance a Pearsonova korelace je základem analýzy hlavních komponent, faktorové analýzy, jakož i dalších vícerozměrných analýz pracujících s lineární závislostí proměnných
- Předpokladem výpočtu kovariance a Pearsonovy korelace je:
  - Normalita dat v obou dimenzích
  - Linearita vztahu proměnných
- Pro vícerozměrné analýzy je nejzávažnějším problémem přítomnost odlehlých hodnot




Lineární vztah – bezproblémové použití Pearsonovy korelace




Korelace je dána dvěma skupinami hodnot – vede k identifikaci skupin objektů v datech




Korelace je dána odlehlou hodnotu – analýza popisuje pouze vliv odlehlé hodnoty

# Analýza kontingenčních tabulek jako princip výpočtu vícerozměrných analýz

- Abundance taxonů (nebo počet jakýchkoliv objektů) na lokalitách lze brát jako kontingenční tabulku a mírou vztahu mezi řádky (lokality) a sloupci (taxony) je velikost chí-kvadrátu

$$\chi^2_{(1)} = \frac{\left[\text{pozorovaná četnost} - \text{očekávaná četnost}\right]^2}{\text{očekávaná četnost}}$$

Počítáno pro každou buňku tabulky

| | ☠ | ☺ |
|---|---|---|
| A | 10 | 0 |
| B | 0 | 10 |

Pozorovaná tabulka

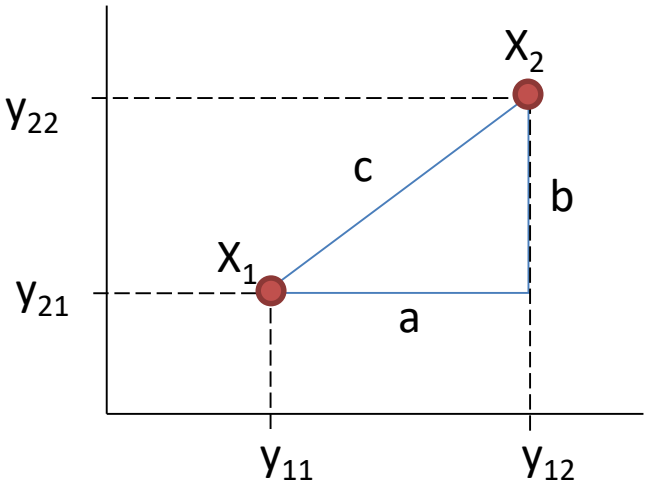
| | ☠ | ☺ |
|---|---|---|
| A | 5 | 5 |
| B | 5 | 5 |

Očekávaná tabulka

Hodnota chí-kvadrátu definuje míru odchylky dané buňky (v našem kontextu vztahu taxon-lokalita) od situace, kdy mezi řádky a sloupci (taxon-lokalita) není žádný vztah

# Euklidovská vzdálenost jako princip výpočtu vícerozměrných analýz

- Nejsnáze představitelným měřítkem vztahu dvou objektů ve vícerozměrném prostoru je jejich vzdálenost
- Nejjednodušším typem této vzdálenosti (bohužel s omezeným použitím na data společenstev) je Euklidovská vzdálenost vycházející z Pythagorovy věty




$$D_1(x_1, x_2) = \sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2}$$

# Základní typy vícerozměrných analýz

## SHLUKOVÁ ANALÝZA

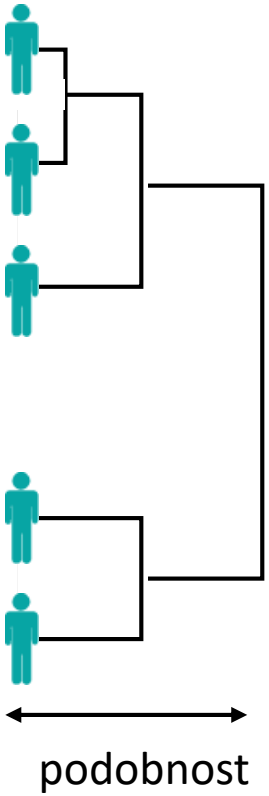
- vytváření shluků objektů na základě jejich podobnosti
- identifikace typů objektů

## ORDINAČNÍ METODY

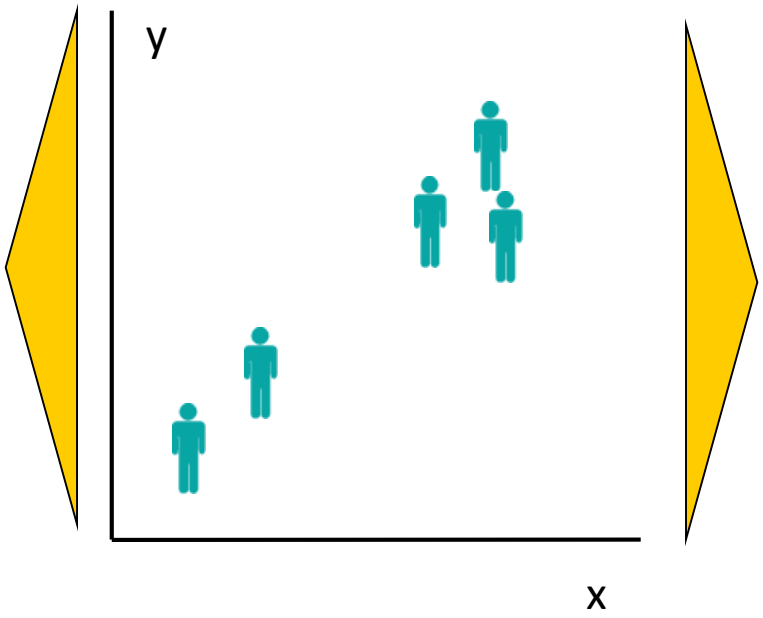
- zjednodušení vícerozměrného problému do menšího počtu rozměrů
- principem je tvorba nových rozměrů, které lépe vyčerpávají variabilitu dat

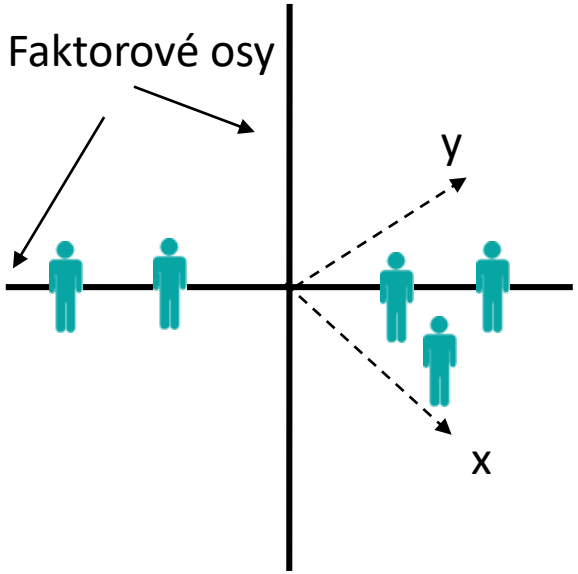
# Typy vícerozměrných analýz

**SHLUKOVÁ ANALÝZA**




**ORDINAČNÍ METODY**

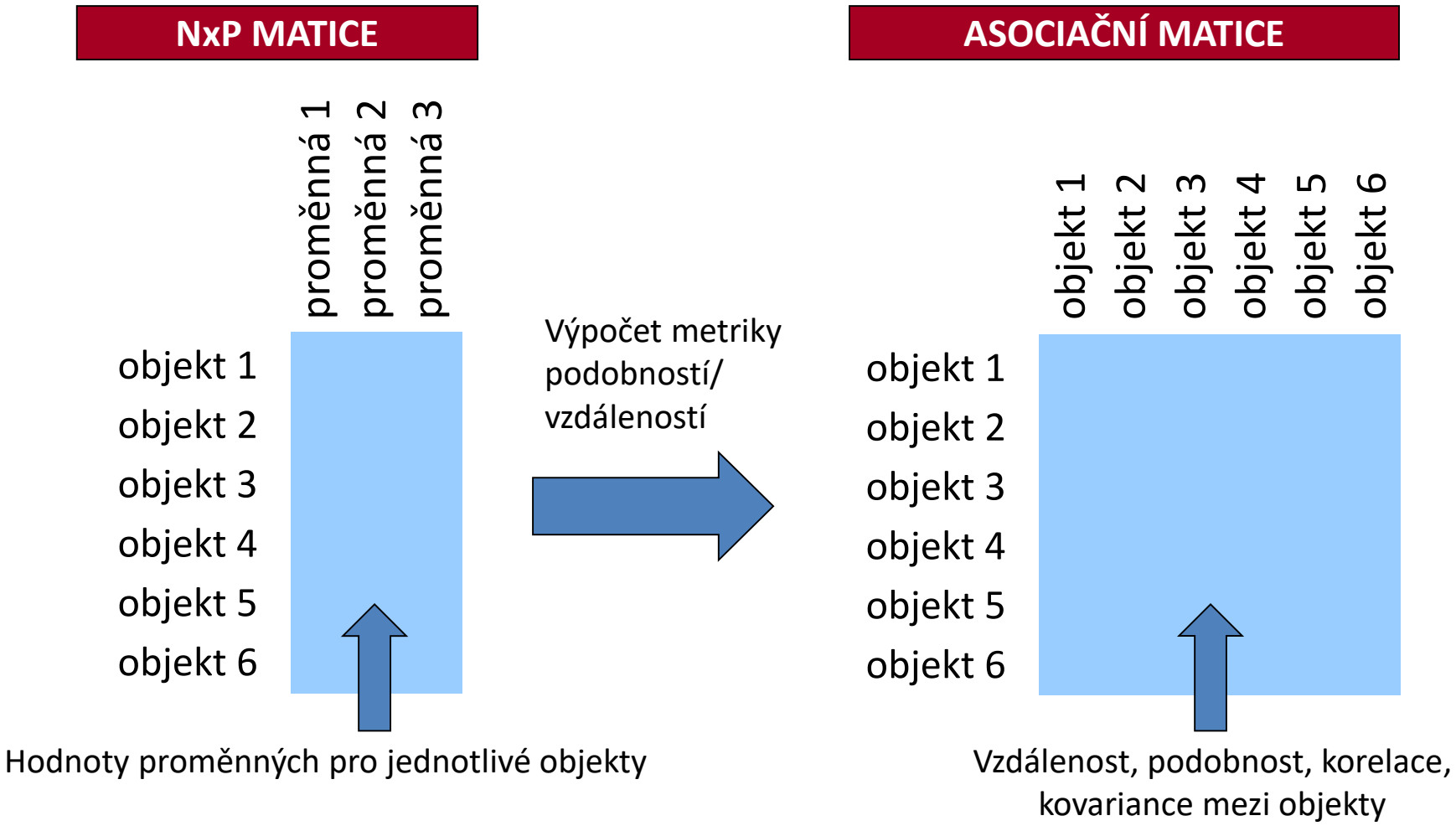


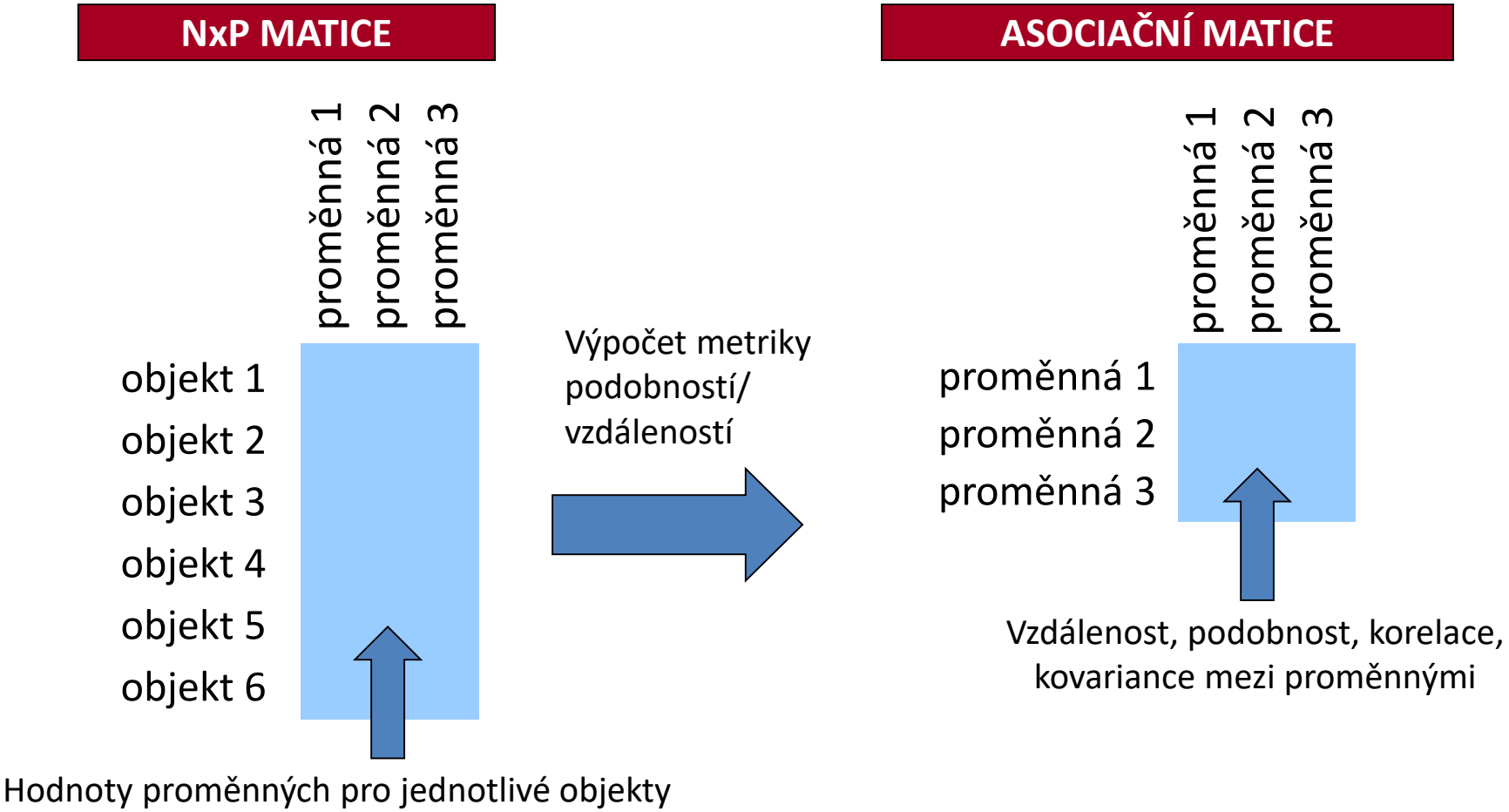
# Pojmy vícerozměrných analýz

- Vícerozměrné metody: Název vícerozměrné vychází z typu vstupních dat, tato data jsou tvořena jednotlivými objekty (i.e. klienti) a každý z nich je charakterizován svými parametry (věk, příjem atd.) a každý z těchto parametrů můžeme považovat za jeden rozměr objektu.
- Maticová algebra: Základem práce s daty a výpočtů vícerozměrných metod je maticová algebra, matice tvoří jak vstupní, tak výstupní data a probíhají na nich výpočty.
- NxP matice: N objektů s P parametry pak vytváří tzv. NxP matici, která je prvním typem vstupu dat do vícerozměrných analýz.
- Asociační matice: Na základě těchto matic jsou počítány matice asociační, na nichž pak probíhají další výpočty; jde o čtvercové matice obsahující informace o podobnosti nebo rozdílnosti (tzv. metriky) buď objektů (Q mode analýza), nebo parametrů (R mode analýza). Měřítko podobnosti se liší podle použité metody a typu dat, některé metody umožňují použití uživatelských metrik.

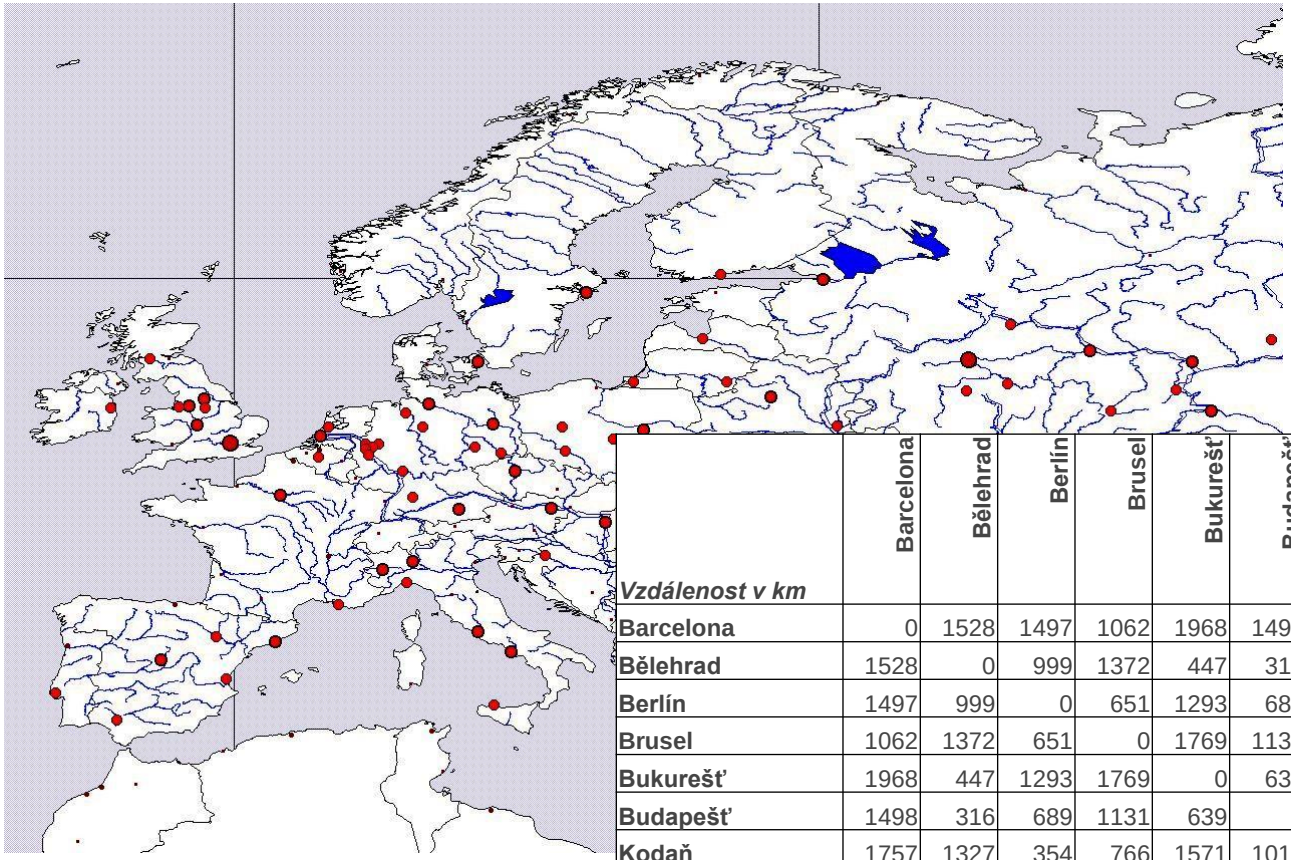
# Asociační matice – Q mode analýza



# Asociační matice – R mode analýza



# Asociační matice – ukázka



Vzdálenost měst v mapě není ničím jiným než maticí vzdálenosti v 2D prostoru

| *Vzdálenost v km* | Barcelona | Bělehrad | Berlín | Brusel | Bukurešť | Budapešť | Kodaň | Dublin | Hamburg | Istanbul | Kiev | Londýn | Madrid |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Barcelona** | 0 | 1528 | 1497 | 1062 | 1968 | 1498 | 1757 | 1469 | 1471 | 2230 | 2391 | 1137 | 504 |
| **Bělehrad** | 1528 | 0 | 999 | 1372 | 447 | 316 | 1327 | 2145 | 1229 | 809 | 976 | 1688 | 2026 |
| **Berlín** | 1497 | 999 | 0 | 651 | 1293 | 689 | 354 | 1315 | 254 | 1735 | 1204 | 929 | 1867 |
| **Brusel** | 1062 | 1372 | 651 | 0 | 1769 | 1131 | 766 | 773 | 489 | 2178 | 1836 | 318 | 1314 |
| **Bukurešť** | 1968 | 447 | 1293 | 1769 | 0 | 639 | 1571 | 2534 | 1544 | 445 | 744 | 2088 | 2469 |
| **Budapešť** | 1498 | 316 | 689 | 1131 | 639 | 0 | 1011 | 1894 | 927 | 1064 | 894 | 1450 | 1975 |
| **Kodaň** | 1757 | 1327 | 354 | 766 | 1571 | 1011 | 0 | 1238 | 287 | 2017 | 1326 | 955 | 2071 |
| **Dublin** | 1469 | 2145 | 1315 | 773 | 2534 | 1894 | 1238 | 0 | 1073 | 2950 | 2513 | 462 | 1449 |
| **Hamburg** | 1471 | 1229 | 254 | 489 | 1544 | 927 | 287 | 1073 | 0 | 1983 | 1440 | 720 | 1785 |
| **Istanbul** | 2230 | 809 | 1735 | 2178 | 445 | 1064 | 2017 | 2950 | 1983 | 0 | 1052 | 2496 | 2734 |
| **Kiev** | 2391 | 976 | 1204 | 1836 | 744 | 894 | 1326 | 2513 | 1440 | 1052 | 0 | 2131 | 2859 |
| **Londýn** | 1137 | 1688 | 929 | 318 | 2088 | 1450 | 955 | 462 | 720 | 2496 | 2131 | 0 | 1263 |
| **Madrid** | 504 | 2026 | 1867 | 1314 | 2469 | 1975 | 2071 | 1449 | 1785 | 2734 | 2859 | 1263 | 0 |

# Vícerozměrné statistické metody

Vícerozměrná data, jejich popis a vizualizace

# Vícerozměrná data

PROMĚNNÉ

OBJEKTY (SUBJEKTY)

| ID | Pohlaví | Věk | Váha | MMSE skóre | Objem hipokampu | … |
|---|---|---|---|---|---|---|
| 1 | muž | 84 | 85,5 | 29 | 7030 | |
| 2 | žena | 25 | 62,0 | 28 | 6984 | |
| 3 | | | | | | |
| 4 | | | | | | |
| … | | | | | | |

Poznámka: proměnné označovány i jako znaky, pozorování, diskriminátory, příznakové proměnné či příznaky

Anglicky označení pouze jedním termínem: feature

# Maticový zápis datového souboru

PROMĚNNÉ

OBJEKTY (SUBJEKTY)

| ID | Pohlaví | Věk | Váha | MMSE skóre | Objem hipokampu | … |
|---|---|---|---|---|---|---|
| 1 | muž | 84 | 85,5 | 29 | 7030 | |
| 2 | žena | 25 | 62,0 | 28 | 6984 | |
| … | | | | | | |

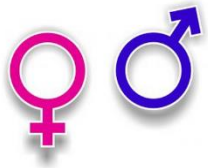
$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

maticový zápis datového souboru *n* objektů (subjektů), které jsou popsané *p* proměnnými

jeden prvek matice $x_{ij}$ je hodnota *j*-té proměnné u *i*-tého objektu (subjektu), přičemž $j = 1, \ldots, p$ a $i = 1, \ldots, n$

# Typy dat - opakování

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



  - Nominální data

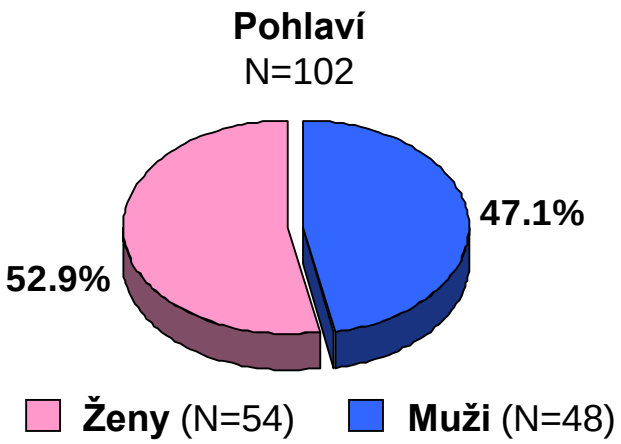
  - Ordinální data

- **Kvantitativní data:**
  - Intervalová data

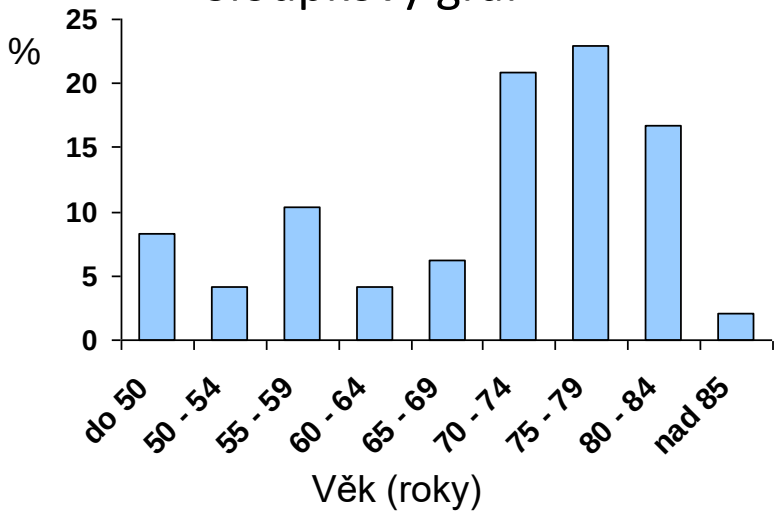

  - Poměrová data

# Vizualizace jednorozměrných dat - opakování

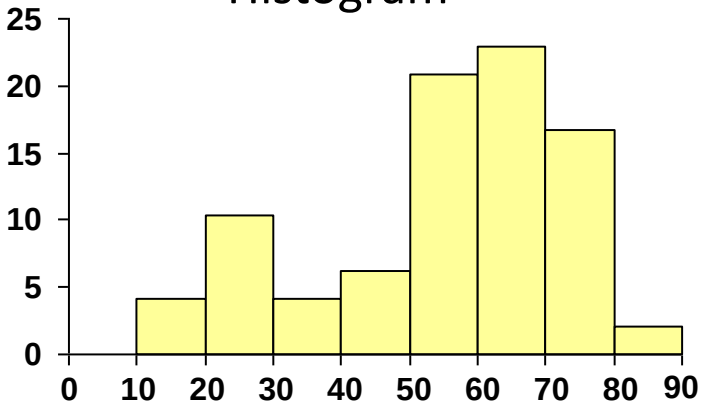
## Koláčový graf



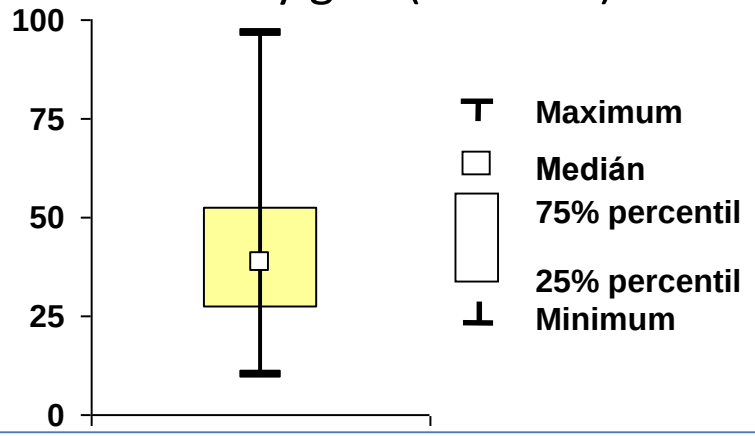

## Sloupkový graf




## Histogram




## Krabicový graf (Box Plot)



# K čemu nám může pomoci vizualizace dat?

→ odhalení problémů v datech

| id | vek | pohlavi | cholesterol | vyska | vaha | obvod_pasu | obvod_boku | BMI | sys_tlak | dia_tlak |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 38 | Z | 4.6 | 164 | 45 | 60 | 87 | 16.7 | 120 | 80 |
| 2 | 36 | Z | 4.35 | 167 | 90 | 97 | 112 | 32.3 | 130 | 80 |
| 3 | 26 | Z | | 178 | 70 | 72 | 94 | 22.1 | 127 | 80 |
| 4 | 25 | Z | 4.2 | 165 | 59 | 65 | 92 | 21.7 | 130 | 80 |
| 5 | 47 | M | 5.65 | 158 | | 92 | 96 | 26.8 | 155 | 90 |
| 6 | 21 | Z | 6.35 | 172 | 61 | 69 | 98 | 20.6 | 135 | 80 |
| 7 | 23 | Z | 3.45 | 170 | 82 | 92 | 113 | 28.4 | 130 | 80 |
| 8 | 35 | M | 7.99 | 179 | 90 | 101 | 110 | 28.1 | 140 | 88 |
| 9 | 33 | Z | 4.88 | 167 | 57 | 70 | 92 | 20.4 | 140 | 85 |
| 10 | 48 | Z | 9.56 | 164 | 70 | 93 | 107 | 26.0 | 250 | 97 |
| 11 | 25 | M | 3.1 | 186 | 75 | 81 | 102 | 21.7 | 120 | 70 |
| 12 | 41 | Z | 10 | 167 | 62 | 71 | 101 | 22.2 | 140 | 90 |
| 13 | 29 | ZZ | 4.2 | 165 | 58 | 66 | 98 | 21.3 | 120 | 80 |
| 14 | 24 | M | 5.62 | 174 | 80 | 92 | 107 | 26.4 | 156 | 90 |
| 15 | 58 | Z | 7.9 | 164 | 63 | 73 | 100 | 23.4 | 135 | 90 |

Chybné hodnoty

Chybějící hodnoty

Odlehlé hodnoty

# Problémy v datech – 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

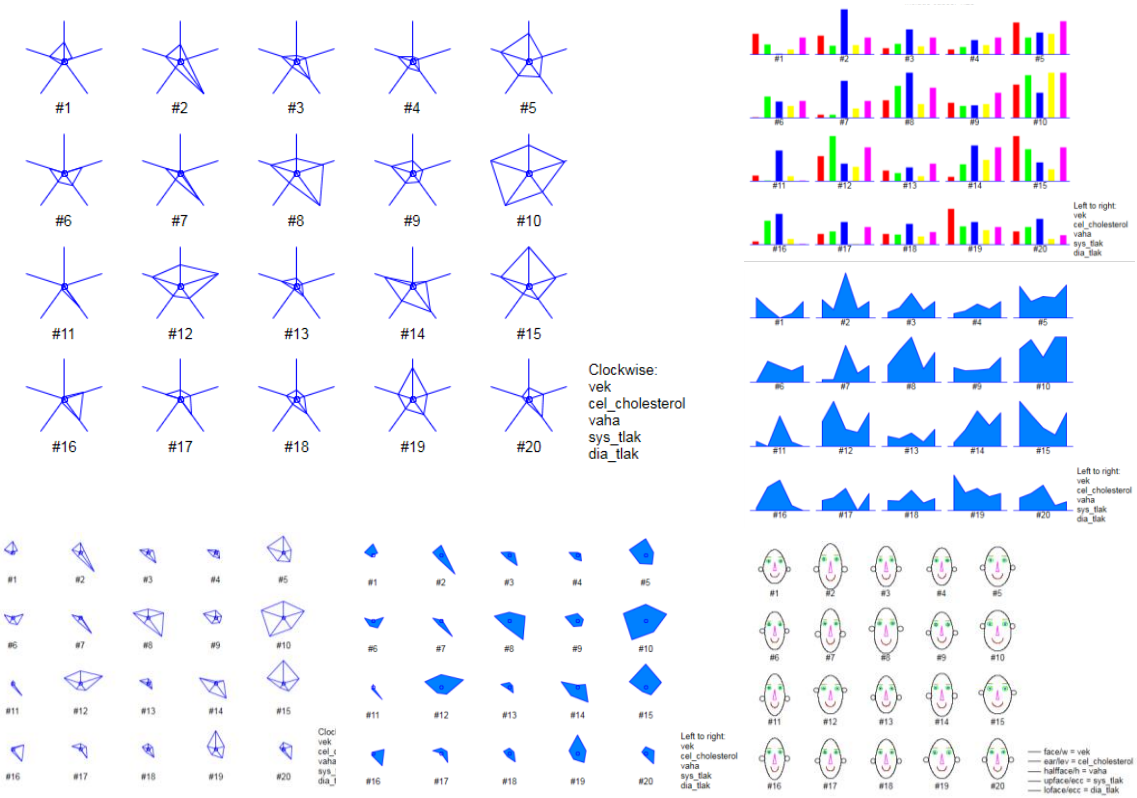
# Problémy v datech – odlehlé hodnoty

- k identifikaci odlehlých hodnot mohou pomoci např. 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

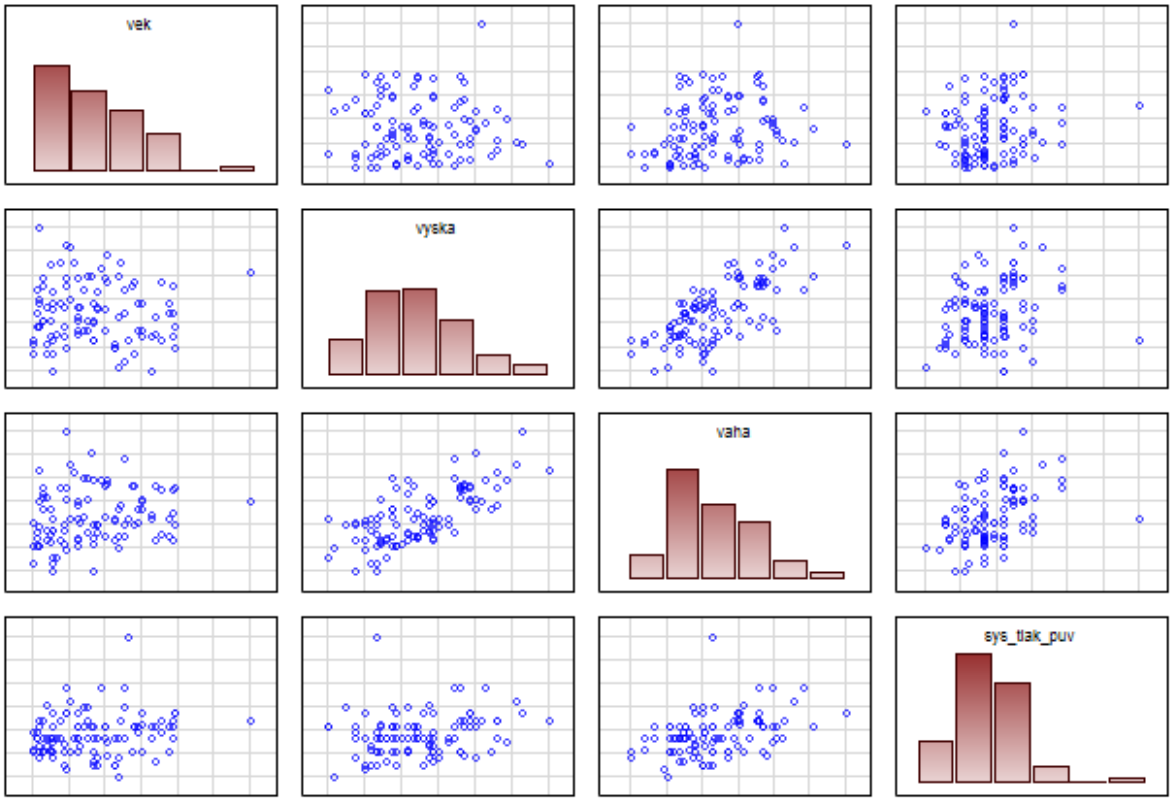
# Vizualizace vícerozměrných dat

- maticové grafy
- krabicové grafy pro více proměnných
- ikonové (symbolové) grafy (v softwaru Statistica: Graphs – Icon Plots...):
  - profilové sloupce
  - profily
  - paprskové (hvězdicové) grafy
  - polygony
  - pavučinové grafy
  - Chernoffovy tváře



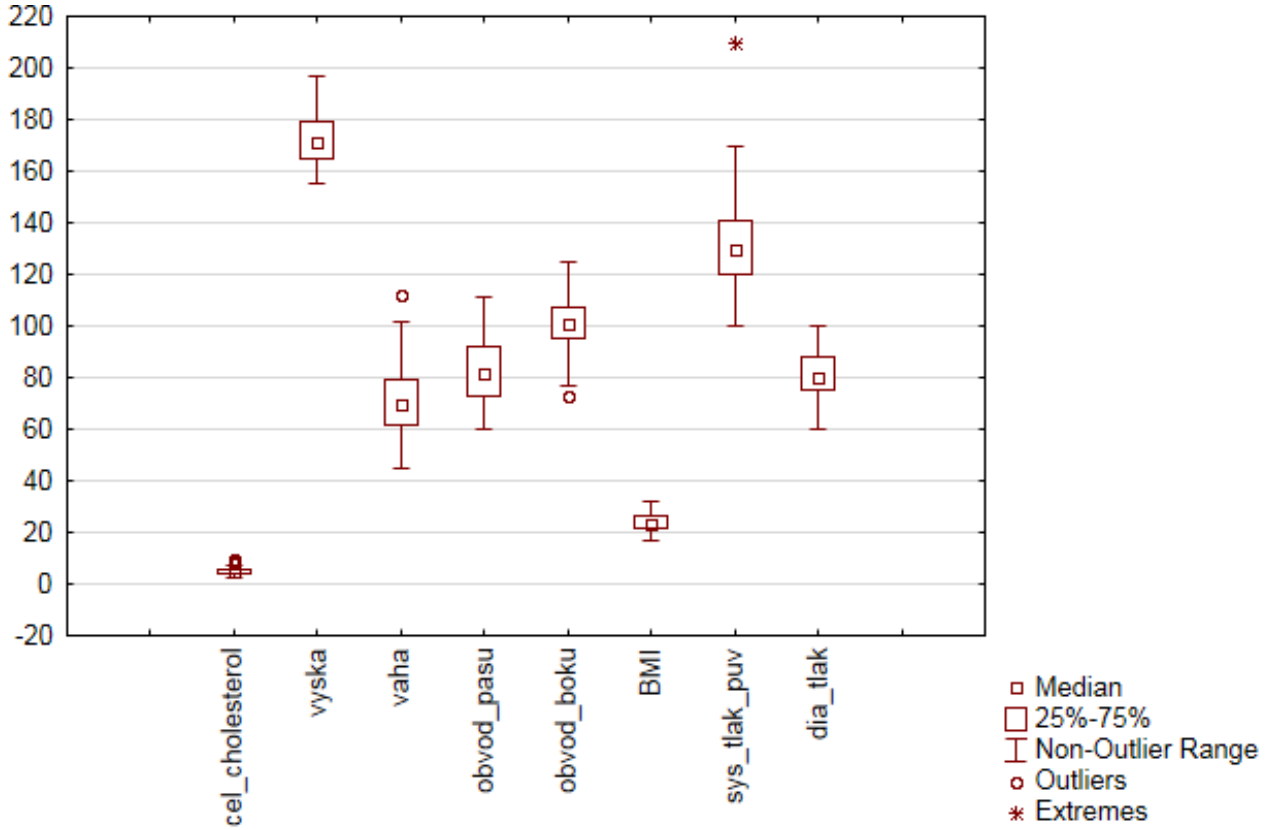
# Maticový graf

- vykreslení vztahu více spojitých proměnných
- v softwaru Statistica: Graphs – Matrix Plots...
- upozornění:  nastavení, jak se vypořádat s chybějícími hodnotami



# Krabicové grafy pro více proměnných

- ukáží nám, zda mají proměnné podobný rozsah hodnot
- v softwaru Statistica: označit příslušné sloupečky v datech – Graphs – Graphs of Block Data – Box Plot: Block columns

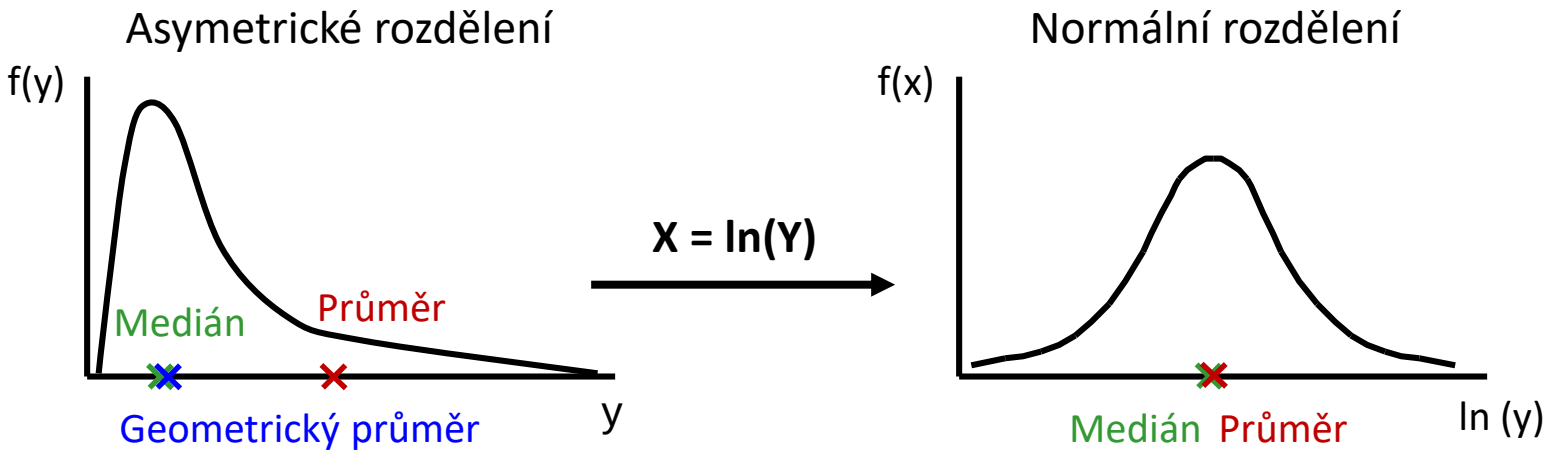


# Typy transformací a jiných úprav vícerozměrných dat

- normalizace dat (= převod na normální rozdělení)
- standardizace dat
- min-max normalizace
- centrování dat
- odstranění vlivu kovariát na jiné proměnné

# Normalizace dat

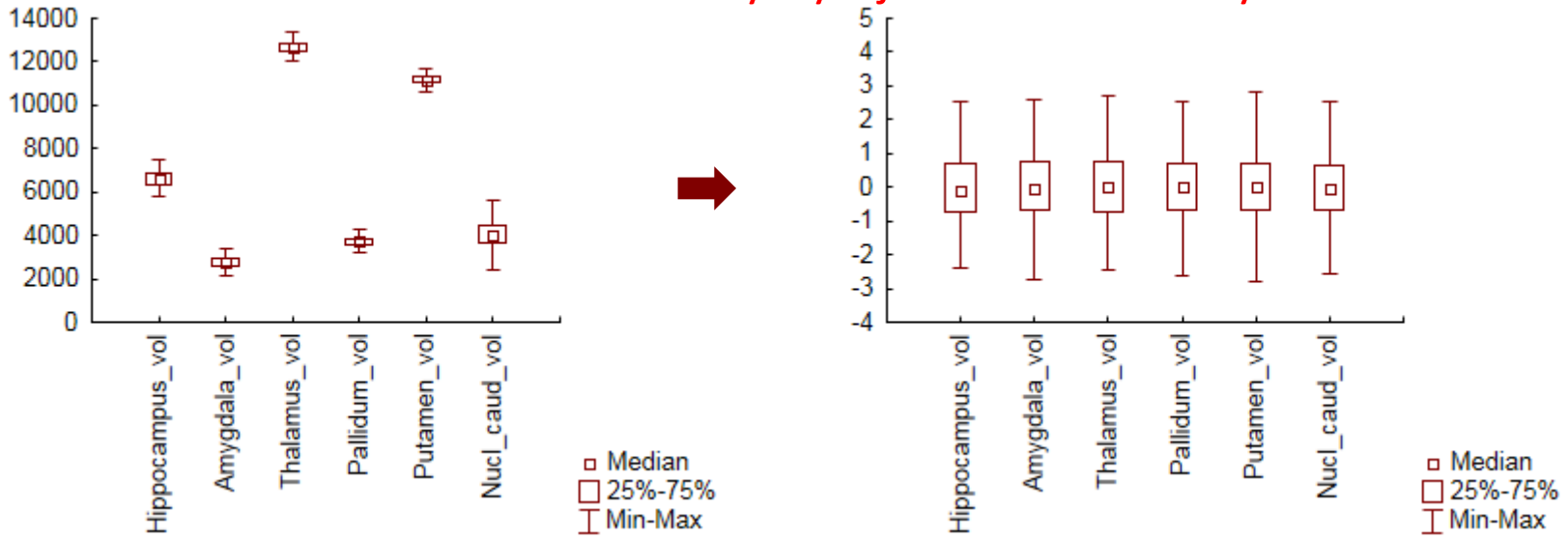
- převod na normální rozdělení (normalita je předpokladem řady statistických testů).
- např. **logaritmická transformace**: X = ln(Y) nebo X = ln(Y+1), pokud data obsahují hodnotu 0




- 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**

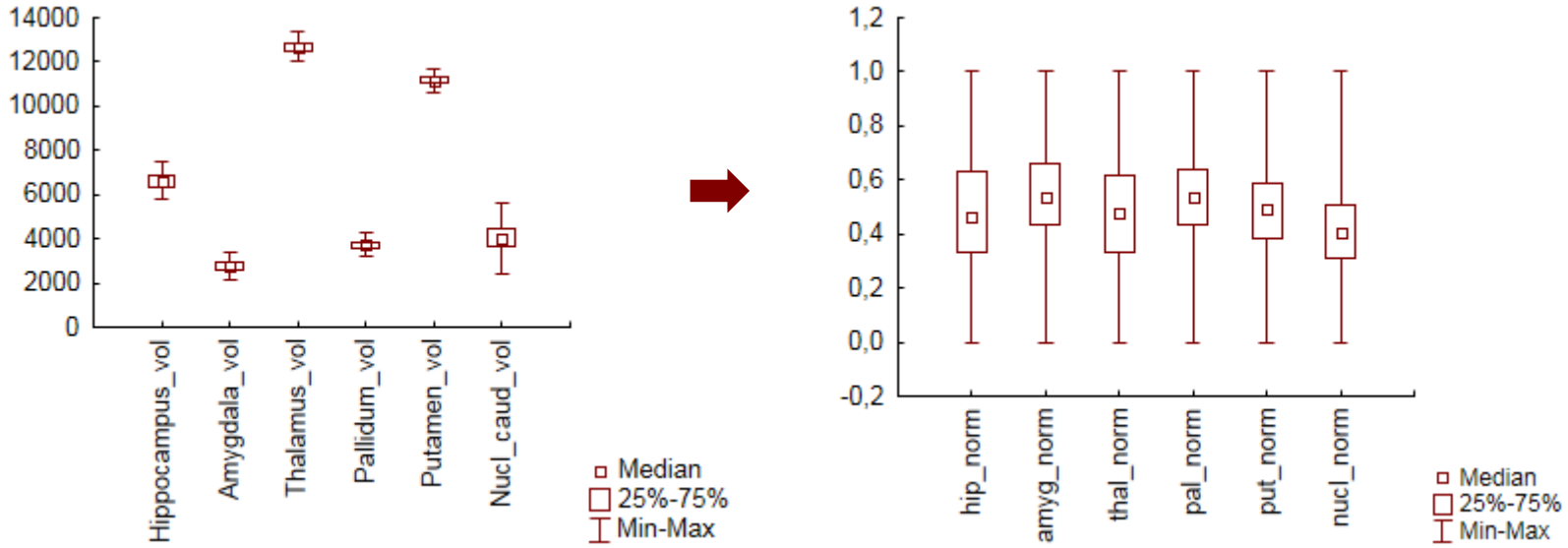
# Standardizace dat

- důvod: převod proměnných na stejné měřítko
- standardizace: $z_i = \frac{x_i - \bar{x}}{s}$ (tzn. odečtení průměru od jednotlivých hodnot a podělení směrodatnou odchylkou)
- proměnné budou mít rozsah přibližně od -3 do 3
- získáme tím současně i tzv. z-skóre (které vyjadřuje, o kolik směrodatných odchylek se i-tá hodnota odchýlila od průměru)
- pozor: standardizace je nevhodná v případě, že proměnné nemají normální rozdělení a že se v datech vyskytují odlehlé hodnoty!!!



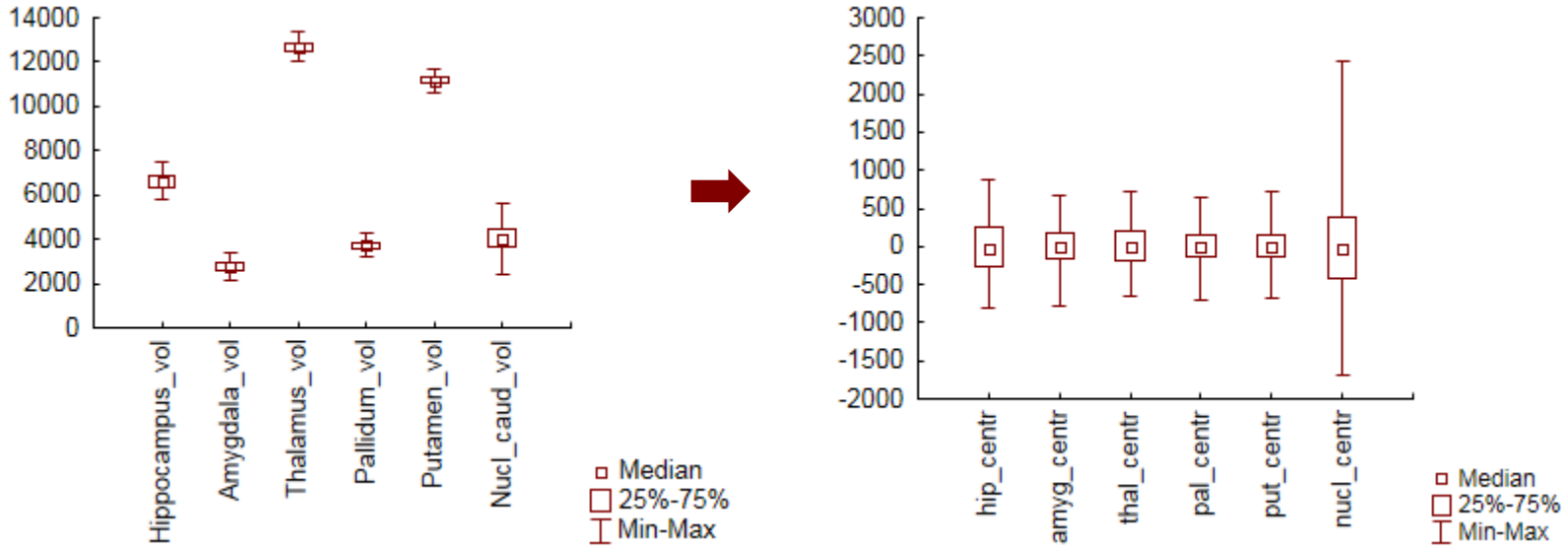
# Min-max normalizace

- důvod: převod proměnných na stejné měřítko
- oproti standardizaci vhodná i na proměnné nemající normální rozdělení či obsahující odlehlé hodnoty
- min-max normalizace: $y_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$
- rozsah hodnot proměnných po min-max normalizaci je od 0 do 1



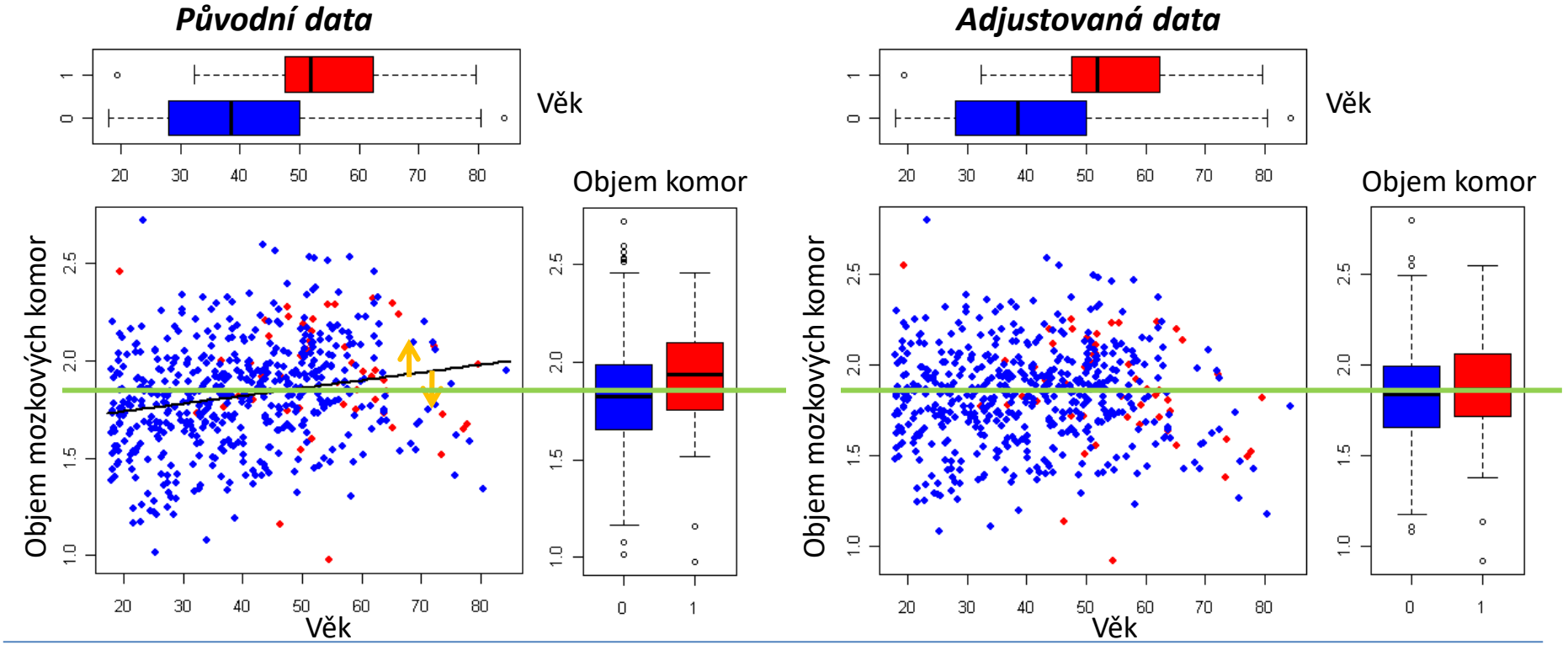
# Centrování dat

- odečtení průměru od dat – získáme novou proměnnou, která bude mít průměr roven nule
- důvod: centrování je důležitou podmínkou některých pokročilých statistických metod (např. klasifikačních)
- centrování: $z_i = x_i - \bar{x}$



# Odstranění vlivu kovariát (tzv. adjustace)

1. V prvním kroku definujeme regresní model vztahu kovariáty (např. věku) a dané proměnné
2. Pro každého pacienta je vypočteno jeho reziduum od regresní přímky ↑↓
3. Reziduum (představující hodnotu parametru po odečtení vlivu věku, jeho průměr je 0) je přičteno k průměrné hodnotě parametru —
4. Výsledná adjustovaná hodnota má odečten vliv věku, ale zároveň není změněna číselná hodnota parametru



# Vícerozměrné statistické metody

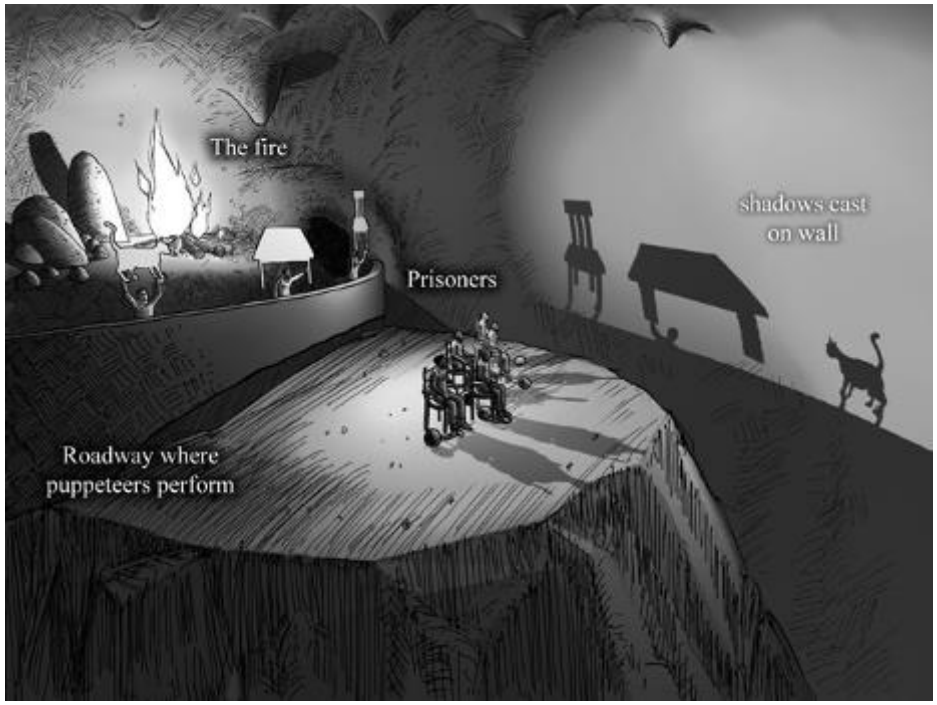
Jednorozměrná  statistická analýza jako předpoklad vícerozměrné analýzy dat

# Význam statistické analýzy dat

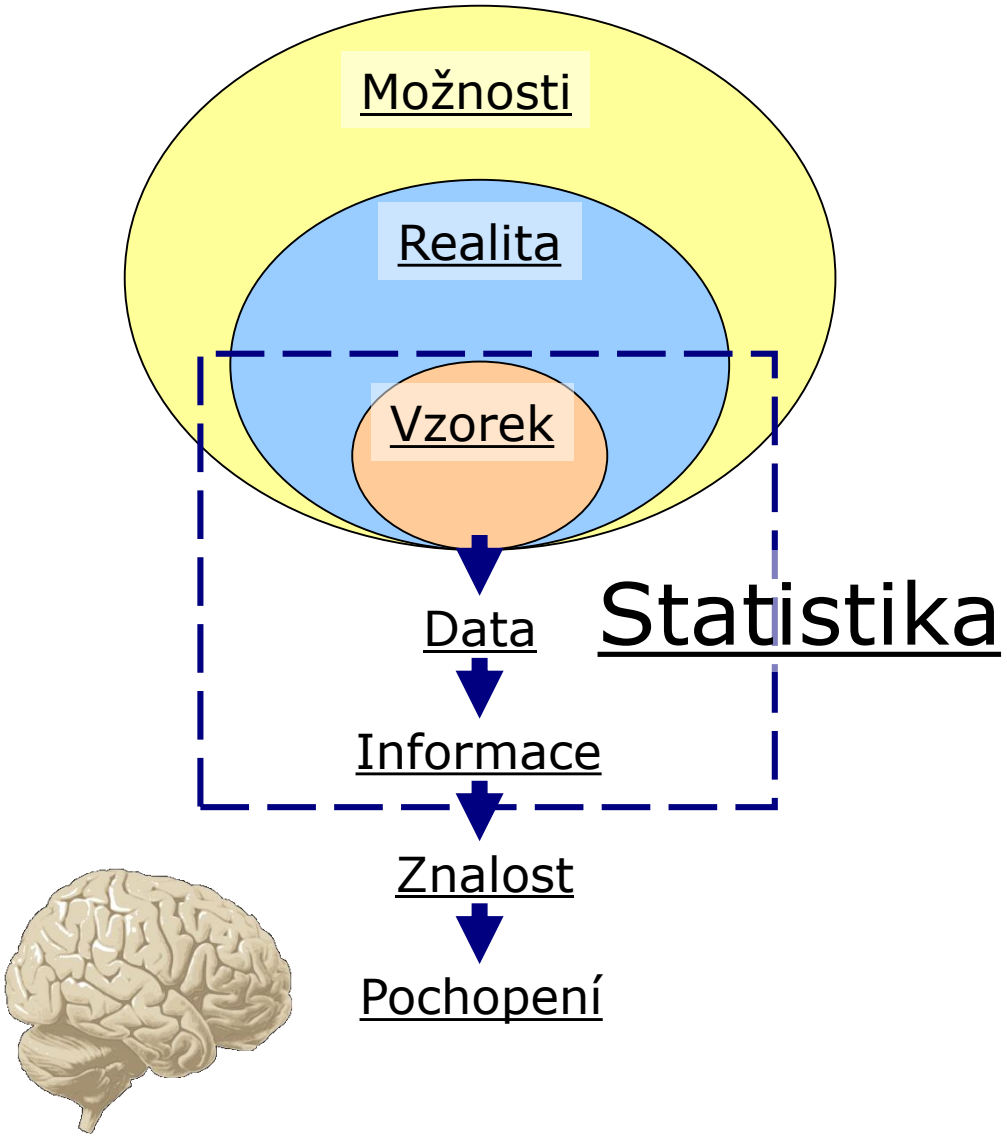
- Výzkum na základě sběru dat je naším způsobem porozumění realitě
- Ale jak přesné a pravdivé je naše porozumění?

Statistika je jedním z nástrojů vnášejících do našich výsledků určitou spolehlivost.

Statistiku můžeme považovat za ekvivalent k mikroskopu či jinému laboratornímu nástroji



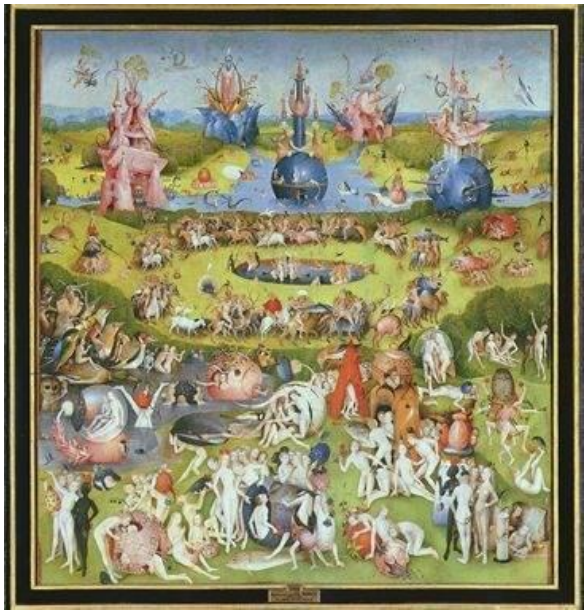
# Co může statistika říci o naší realitě?




- Statistika není schopna činit závěry o jevech neobsažených v našem vzorku.
- Statistika je nasazena v procesu získání informací z vzorkovaných dat a je podporou v získání naší znalosti a pochopení problému.
- Statistika není náhradou naší inteligence !!!

# Variabilita jako základní pojem ve statistice

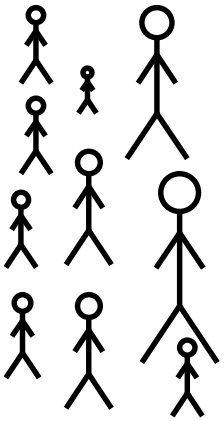
- Naše realita je variabilní a statistika je vědou zabývající se variabilitou
- Korektní analýza variabilita a její pochopení přináší užitečné informace o naší realitě
- V případě deterministického světa by statistická analýza nebyla potřebná



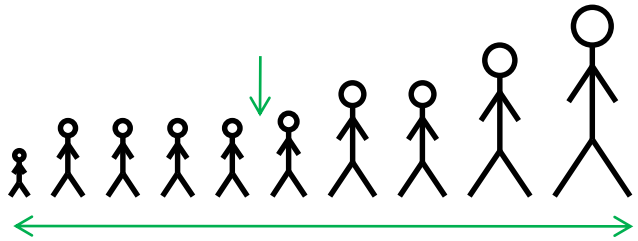
# Práce s variabilitou v analýze dat

- V analýze dat existují dva hlavní přístupy k práci s variabilitou

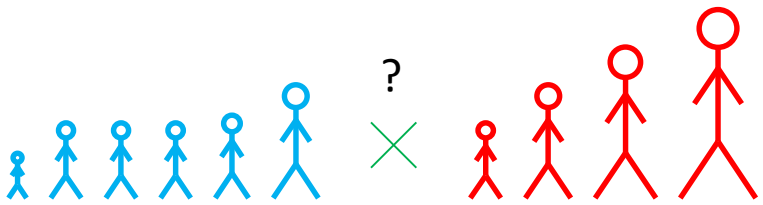
Variabilita dat



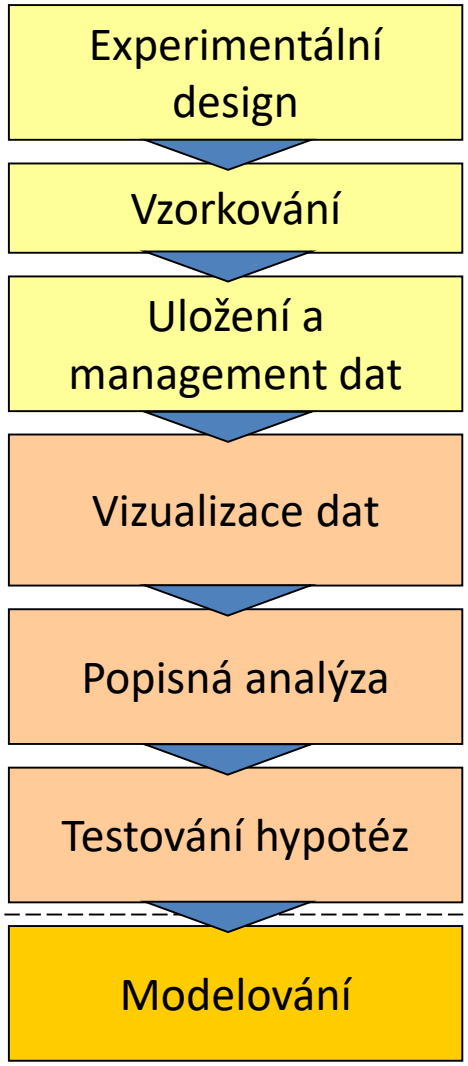
Popisná analýza: charakterizace variability



Testování hypotéz: vysvětlení variability



# Obecné schéma aplikace statistické analýzy




Jak velký vzorek je nezbytný pro statisticky relevantní výsledky?
Klíčová stratifikační kritéria cílové populace.

Vzorkovací plán zabezpečující náhodnost a reprezentativnost vzorku.

Uložení dat ve vhodné formě a jejich vyčištění předcházející vlastní analýze je klíčovým krokem statistické analýzy.

Grafická inspekce dat je nezbytným krokem analýzy vzhledem ke schopnosti lidského mozku primárně akceptovat obrazová data. Poskytne vhled do dat, představu o jejich rozložení, vazbách proměnných apod.

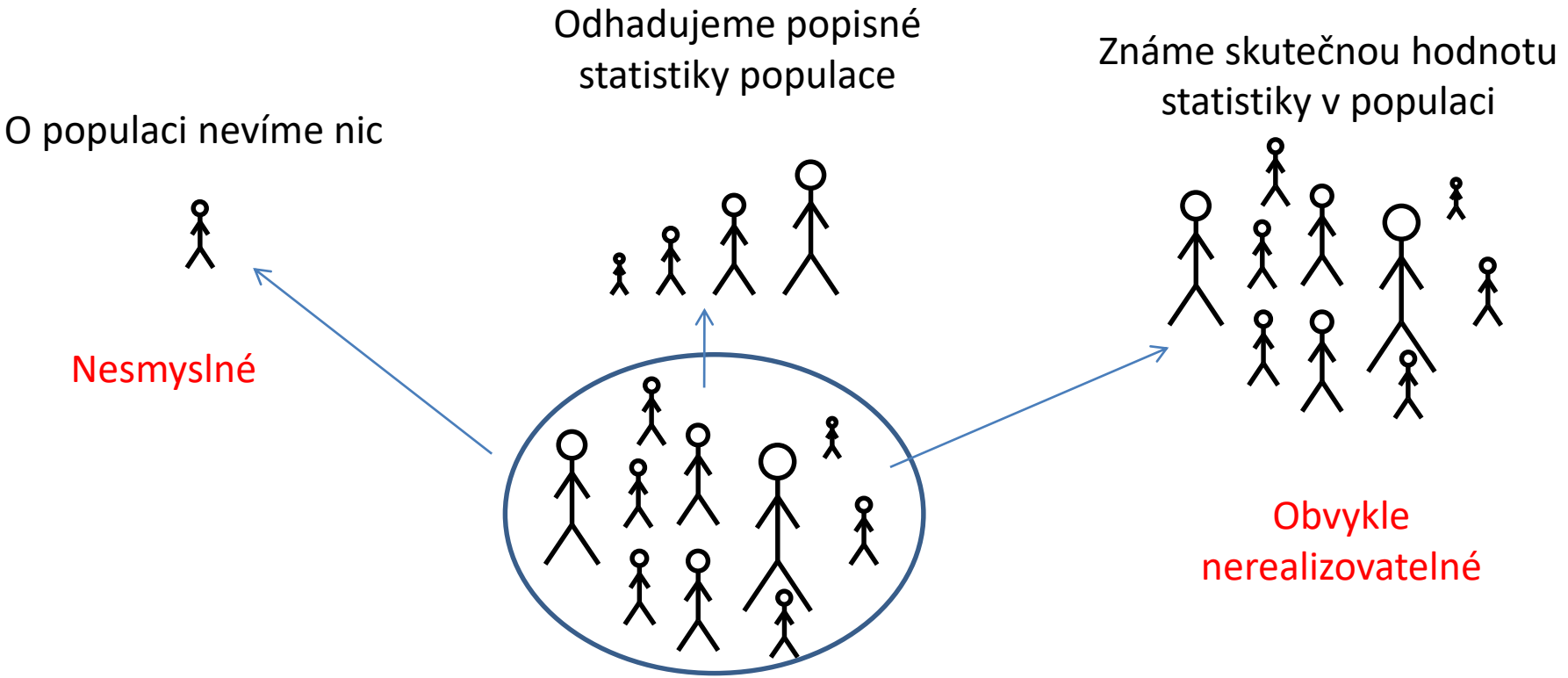
Popisná analýza umožňuje vyhodnotit srovnáním s existující literaturou realističnost naměřených rozsahů dat.

Testování vazeb mezi různými proměnnými s cílem navzájem vysvětlit jejich variabilitu a tím přispět k pochopení řešeného problému.

Možným vyvrcholením analýzy je využití získaných znalostí a pochopení problému k vytvoření prediktivních modelů.

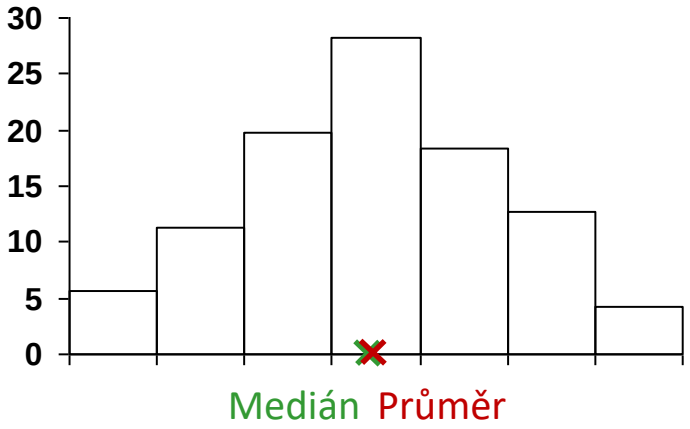
# Popisná statistika: odhad reality

- Při výpočtu popisné statistiky počítáme popisnou statistiku vzorku, která je zároveň odhadem pro celou cílovou populaci
- Skutečnou hodnotu statistiky v cílové populaci nemůžeme poznat bez vzorkování celé cílové populace



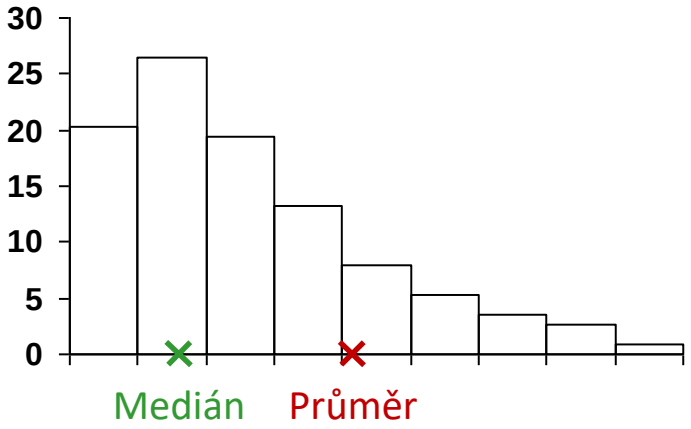
# Popis kvantitativních dat

## Symetrická data




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

## Asymetrická data

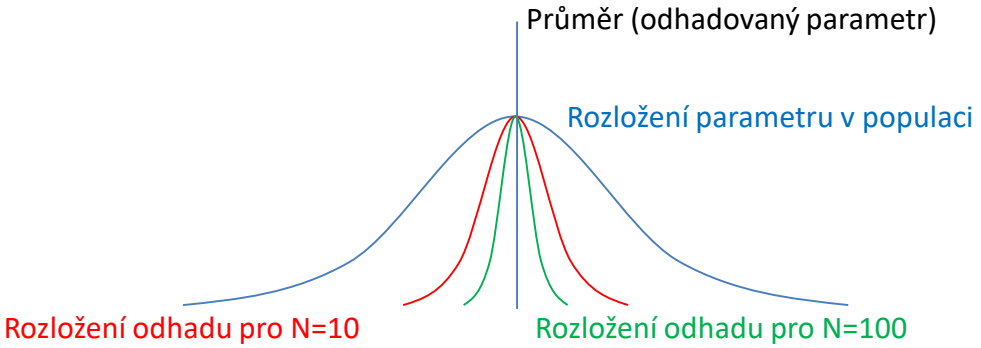



| | MMSE |
|---|---|
| N | 833 |
| Medián (Median) | 27 |
| Minimum | 18 |
| Maximum | 30 |

# Koncept intervalu spolehlivosti a jeho interpretace

- Při výpočtu odhadu popisné statistiky nás zajímá nejenom její vlastní hodnota (bodový odhad) ale také její rozsah spolehlivosti

- Interval spolehlivosti závisí na:
  - Velikosti vzorku
  - Variabilitě dat
  - Požadované spolehlivosti



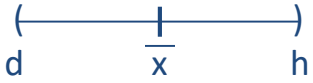

- Interval spolehlivosti lze spočítat pro jakoukoliv statistiku (průměr, směrodatná odchylka, korelace, procentuální zastoupení apod.)
- Interval spolehlivosti poskytuje vodítko jak „spolehlivé“ jsou naše výsledky a s jakou pravděpodobností jich je možné opakovaně dosáhnout
- 95% interval spolehlivosti je rozsah hodnot, do nějž se při opakování studie trefíme s 95% pravděpodobností
- Tvrzení, že v rozsahu 95% intervalu spolehlivosti leží s 95% pravděpodobností skutečný průměr populace není pravdivé, skutečný průměr populace neznáme !!!

# Ovlivnění šířky intervalu spolehlivosti

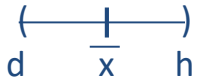
**Interval spolehlivosti je tedy UŽŠÍ s**:

- **větším N** (větší velikostí vzorku)

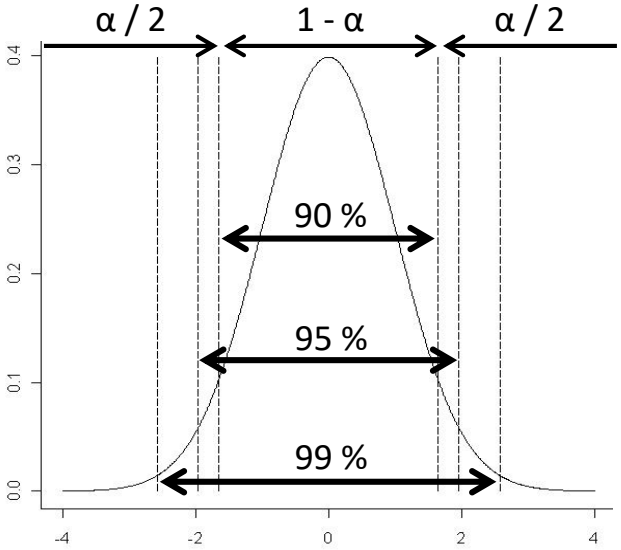
n=20:




n=100:




- menší variabilitou dat
- **menší spolehlivostí**



# Statistické testování – základní pojmy

- Nulová hypotéza $H_O$ — $H_O$: sledovaný efekt je nulový
- Alternativní hypotéza $H_A$ — $H_A$: sledovaný efekt je různý mezi skupinami
- Testová statistika

$$\text{Testová statistika} = \frac{\text{Pozorovaná hodnota} - \text{Očekávaná hodnota}}{\text{Variabilita dat}} * \sqrt{\text{Velikost vzorku}}$$

- Statistická významnost (p) – odvozena z testové statistiky a znamená pravděpodobnost, že pozorovaný rozdíl je výsledkem pouhé náhody

# p-hodnota („p-value“, „p-level“)

- Neboli **dosažená hladina významnosti testu**.
- Značka: *p*
- Je to pravděpodobnost, s jakou bychom mohli obdržet pozorovaná data nebo data stejně, či ještě více odporující nulové hypotéze, za předpokladu, že je nulová hypotéza pravdivá.
- Čím menší je *p*, tím neudržitelnější čili méně důvěryhodná je nulová hypotéza.

- Hodnocení, kdy je výsledek testu statisticky významný:
  - Máme zvolenu hladinu významnosti testu (např. $\alpha$=0,05).
  - Dvě možné situace:
    1. $p < \alpha$ – **zamítáme $H_0$** – statisticky významný výsledek testu
    2. $p \geq \alpha$ – **nezamítáme $H_0$**

# Důležité poznámky k testování hypotéz

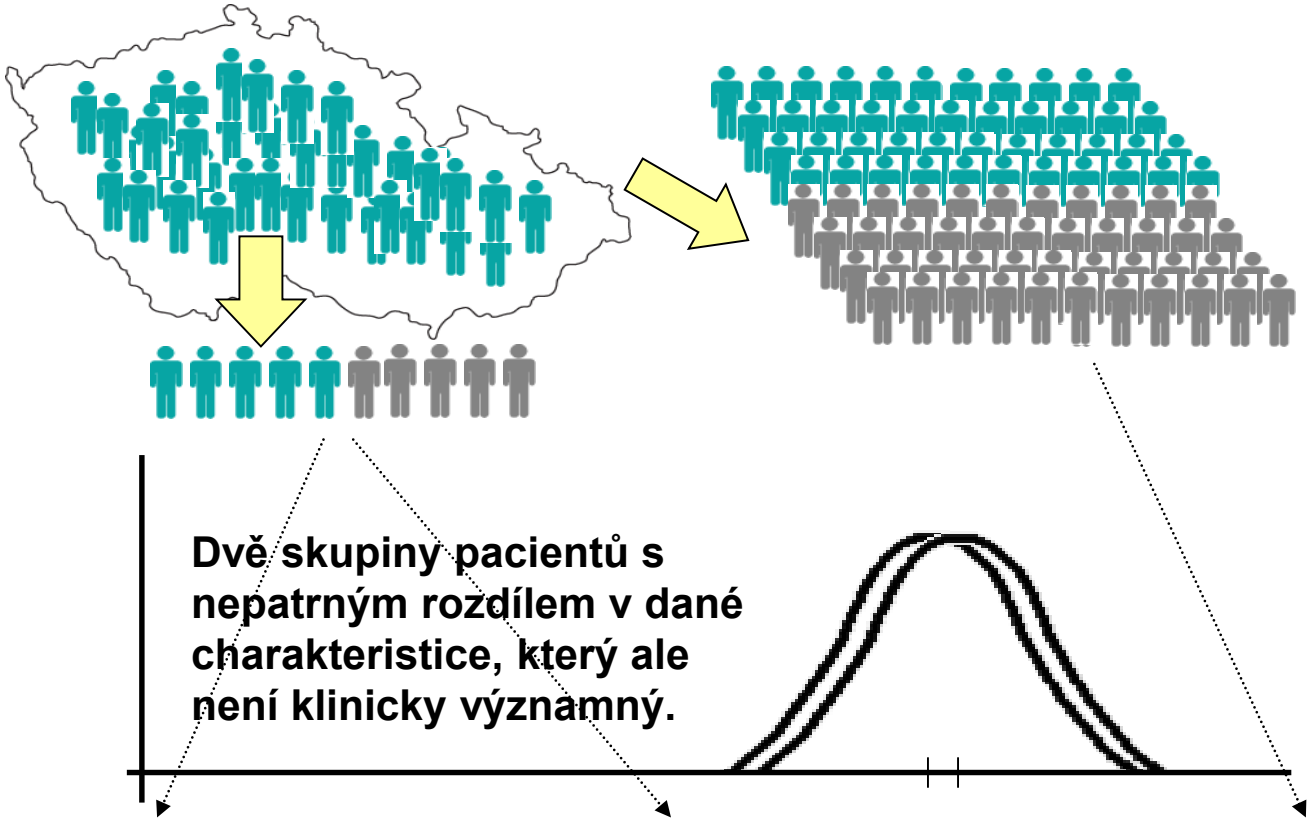
- **Nezamítnutí nulové hypotézy neznamená automaticky její přijetí!** Může se jednat o situaci, kdy pro zamítnutí nulové hypotézy nemáme dostatečné množství informace.
- **Dosažená hladina významnosti testu** (ať už 0,05, 0,01 nebo 0,10) **nesmí být slepě brána jako hranice pro existenci/neexistenci testovaného efektu.** Neexistuje jasná hranice pro významnost či nevýznamnost – často je velmi malý rozdíl mezi p-hodnotou 0,04 a p-hodnotou 0,06.
- **Malá p-hodnota nemusí znamenat velký efekt.** Hodnota testové statistiky a odpovídající p-hodnota může být ovlivněna velkou velikostí vzorku a malou variabilitou pozorovaných dat.
- **Výsledky testování musí být nahlíženy kriticky** – jedná se o závěr založený „pouze“ na jednom výběrovém souboru.

# Klinická a statistická významnost

- Samotná statistická významnost nemá žádný reálný význam, je pouze měřítkem náhodnosti hodnoceného jevu
- Pro vyhodnocení reálné významnosti je nezbytné znát i reálně významné hodnoty

| | | Praktická významnost | |
|---|---|---|---|
| | | **ANO** | **NE** |
| Statistická významnost | **ANO** | OK, praktická i statistická významnost je ve shodě, jednoznačný závěr | Významný výsledek je statistický artefakt velkého vzorku, prakticky nevyužitelné |
| | **NE** | Výsledek může být pouhá náhoda, neprůkazný výsledek | OK, praktická i statistická významnost je ve shodě, jednoznačný závěr |

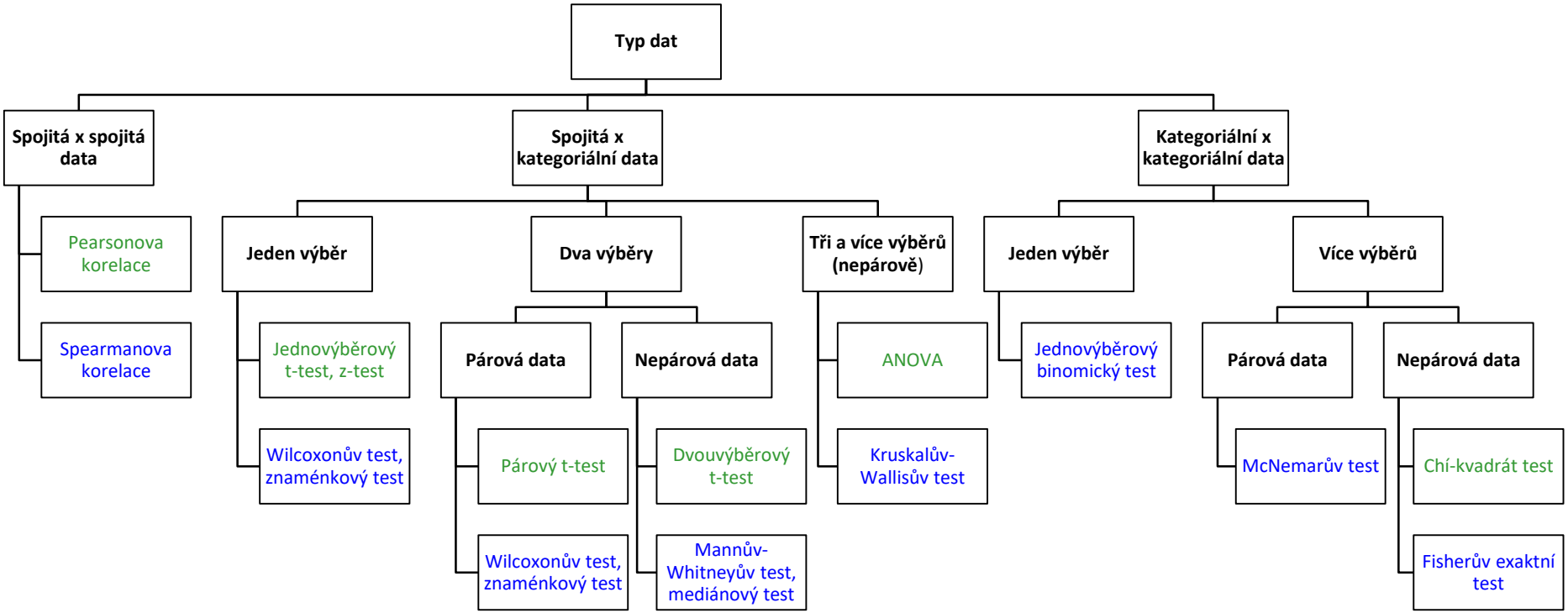
# Vliv velikosti vzorku na výsledky testování




| $n_1 = 10, n_2 = 10$ | $n_1 = 100, n_2 = 100$ | $n_1 = 1000, n_2 = 1000$ |
|---|---|---|
| $p = 0{,}797$ | $p = 0{,}140$ | $p < 0{,}001$ |

**Statistická významnost způsobená velkým N**

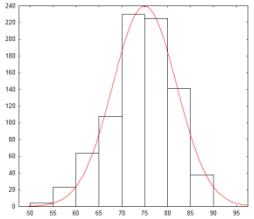
# Základní rozhodování o výběru statistických testů



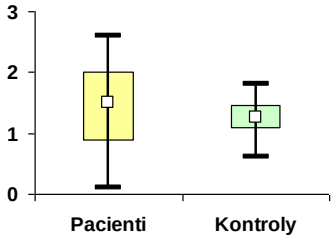

**Poznámka:**
zeleně označeny parametrické testy, u nichž je nutno ověřovat předpoklady
modře označeny neparametrické testy

# Výběr statistického testu se provádí na základě

- **Typu dat** –ordinální, nominální data, nebo spojité hodnoty?
- **Rozdělení** dat – u **parametrických** testů.
  - **Normalita** předpokladem mnoha testů



- **Homogenity rozptylu** srovnávaných skupin – tzn. předpokladu, že rozptyl ve skupinách je přibližně stejný.
  - mnoho testů vyžaduje homogenitu rozptylu




- **Typu hypotézy (srovnání):**
  - 1 skupina vs referenční hodnota (jednovýběrový test)
  - 1 skupina před a po (párový test)
  - 2 skupiny mezi sebou (dvouvýběrový test)
  - Více skupin mezi sebou
- **Typu alternativní hypotézy:** oboustranná vs jednostranná

# Předpoklady statistického testu

- Výše uvedené podmínky pro výběr statistického testu jsou zároveň **předpoklady použití statistického testu**

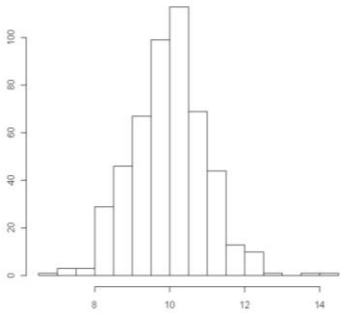
- Další předpoklad: **vyrovnané počty subjektů** ve srovnávaných skupinách –aby byly odhady ve srovnávaných skupinách podobně přesné a spolehlivé

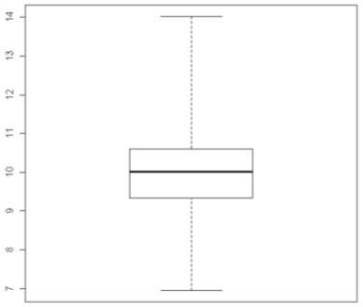
- Splnění **všech předpokladů** je důležité pro **použití statistického testu**

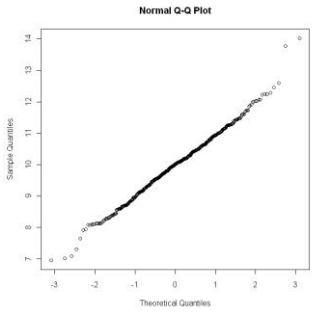
V případě, že tyto předpoklady nejsou splněny, nemůžeme důvěřovat výsledkům testu !!!

# Ověření normality dat

- **Graficky:**
  - histogram
  - krabicový graf (box-plot)
  - Q-Q graf







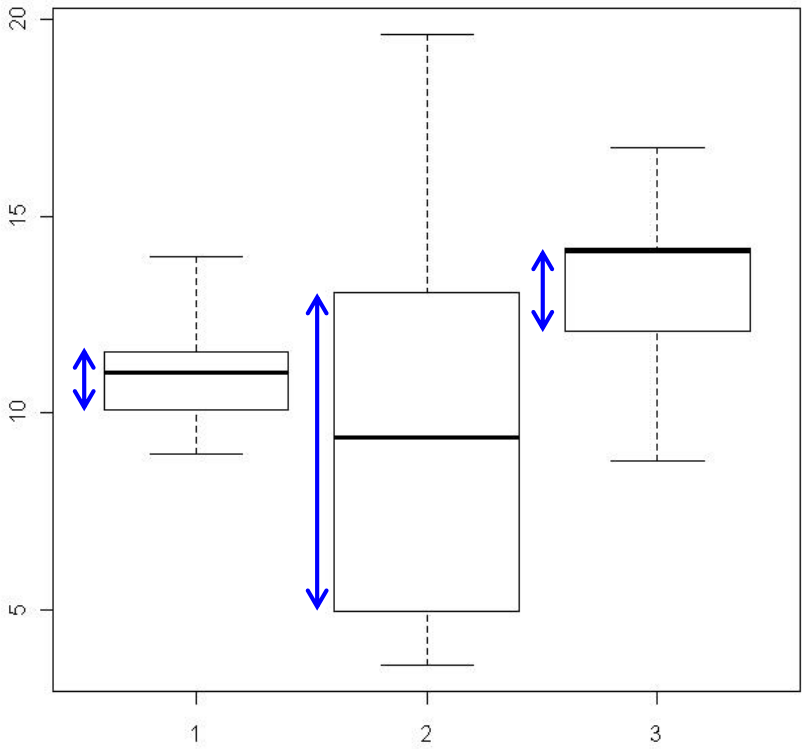

- **Testy normality:**
  - Shapirův-Wilkův test
  - Kolmogorovův-Smirnovův test

- Testy nejsou vždy nejlepším nástrojem! Vždy je důležité se podívat i očima!

- Pokud o sledované veličině prokazatelně víme, že v cílové populaci nabývá normální rozdělení (např. výška lidské postavy), ale v daném souboru normální rozdělení nepotvrdíme, **pak s naším náhodným výběrem není něco v pořádku** – např. není reprezentativní.

# Ověření shody (homogenity) rozptylů

- **Grafické ověření** – krabicový graf, histogram.
- **F-test** (testování shody rozptylů dvou vzorků)
- **Leveneův test** – často používaný (testování shody rozptylů dvou a více vzorků)
- **Bartlettův test**



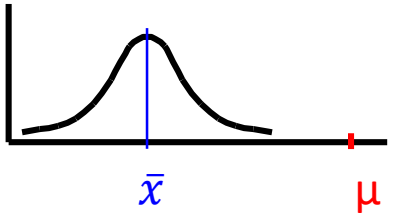
# Parametrické a neparametrické testy

- **Parametrické testy:**
  - Mají předpoklady o rozdělení vstupních dat (např. předpoklad normálního rozdělení), protože se zabývají testováním tvrzení o neznámých parametrech rozdělení (např. střední hodnoty)
  - Mají větší sílu než neparametrické testy

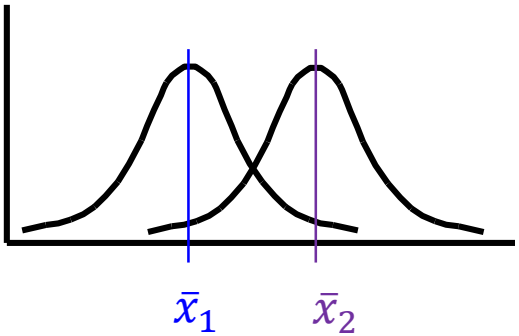
- **Neparametrické testy:**
  - Nemají předpoklady o rozdělení vstupních dat
  - Možné je použít při **asymetrickém** rozdělení nebo **odlehlých** hodnotách
  - Nevyužívají původní hodnoty, ale nejčastěji pouze jejich pořadí – tím dochází **k redukci informační hodnoty** původních dat, a proto mají **menší sílu**
  - Menší sílu testu je možné vykompenzovat větší velikostí vzorku
  - Používání neparametrických testů je „bezpečnější“

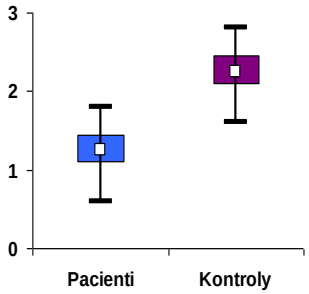
# Jednovýběrové a dvouvýběrové testy

- **Jednovýběrové testy:**
  - **Srovnávají jeden vzorek s referenční hodnotou** (popřípadě se statistickým parametrem cílové populace)
  - Průměrný objem hipokampu u 406 pacientů s MCI v našem souboru vs 6575 $mm^3$ zjištěným při populačním epidemiologickém průzkumu.



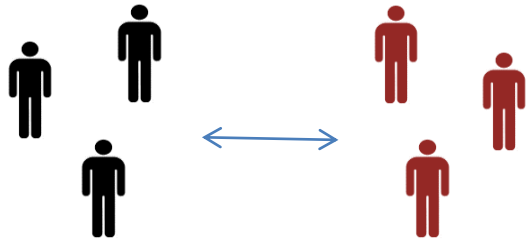

- **Dvouvýběrové testy:**
  - **Srovnáváme dvě skupiny dat**
  - Příklady: srovnání objem hipokampu u mužů a u žen, srovnání kognitivního výkonu podle dvou kategorií věku.



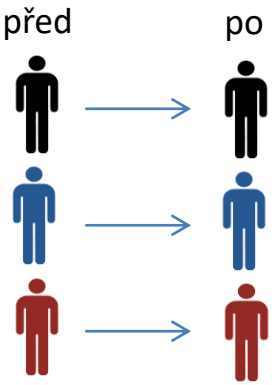



# Párové a nepárové testy

- **Nepárové testy:**
  - **Srovnáváme dvě skupiny dat**, které jsou na sobě **nezávislé** – mezi objekty neexistuje vazba.
  - Příklady: srovnání objem hipokampu u mužů a u žen, srovnání kognitivního výkonu podle dvou kategorií věku.

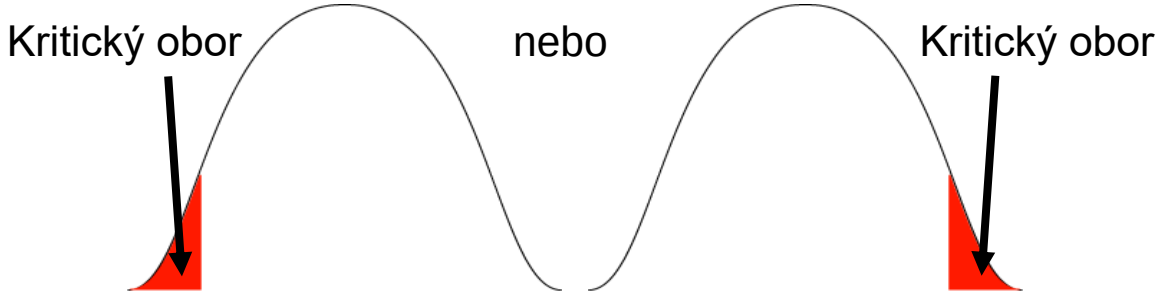


- **Párové testy:**
  - **Srovnáváme dvě skupiny dat**, které jsou na sobě **závislé** – mezi objekty existuje vazba.
  - Příklady: hodnota krevního tlaku **před** začátkem léčby a **po** ukončení léčby

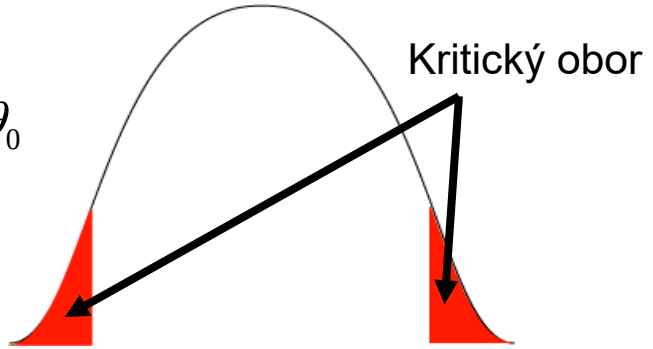


# Jednostranné a oboustranné testy

- **Jednostranné („One-Tailed“) testy:**
  - Jednostranná alternativní hyp.: $H_1 : \theta < \theta_0$ $H_1 : \theta > \theta_0$
  - Např. testujeme, zda je objem mozkové struktury menší u žen než u mužů či zda je průměrná spotřeba tišících léků větší u pacientů než populační průměr apod.



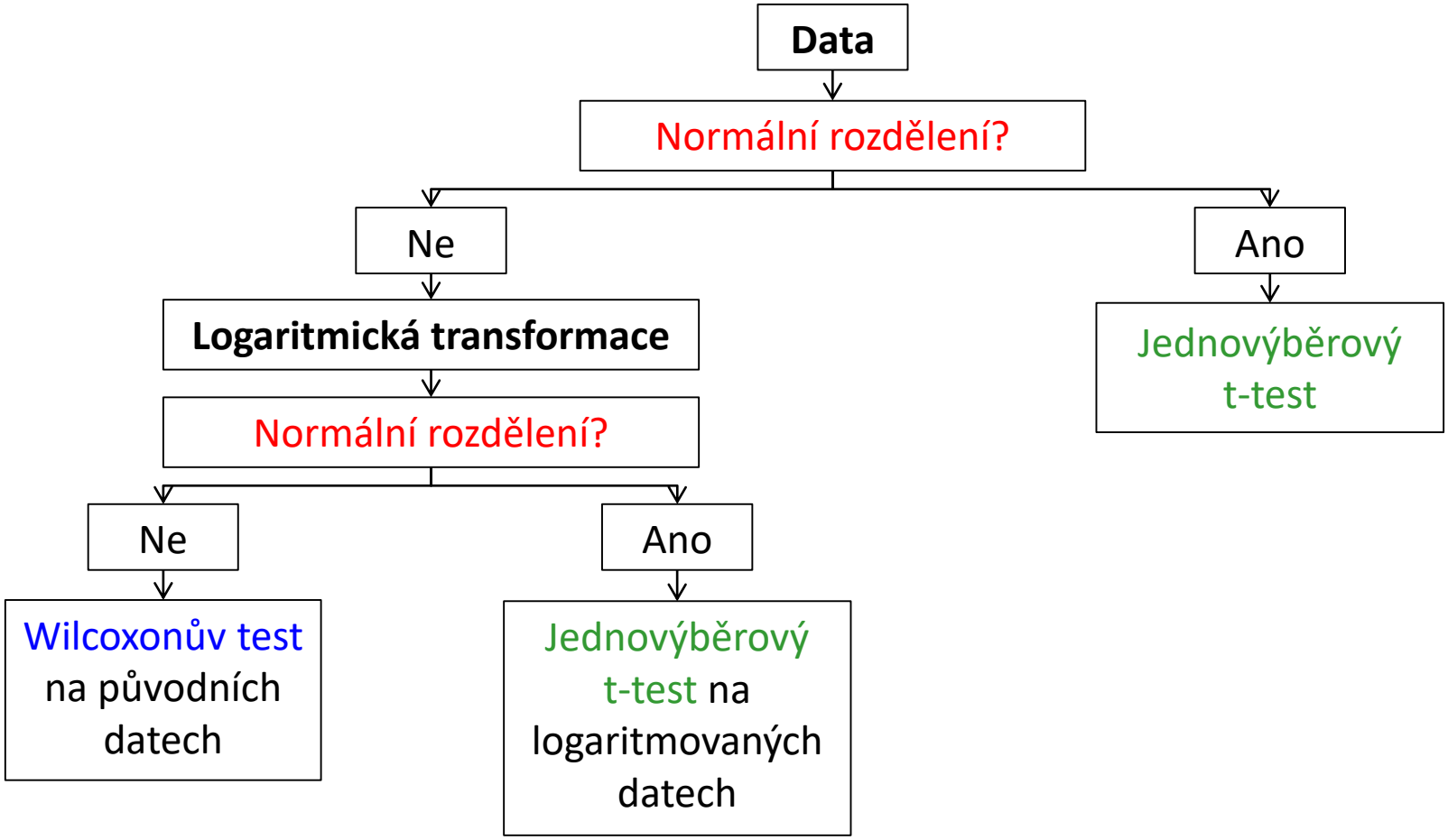

- **Oboustranné („Two-Tailed“) testy:**
  - Oboustranná alternativní hyp.: $H_1 : \theta \neq \theta_0$
  - Např. testujeme, zda se objem mozkové struktury liší u žen a mužů apod.



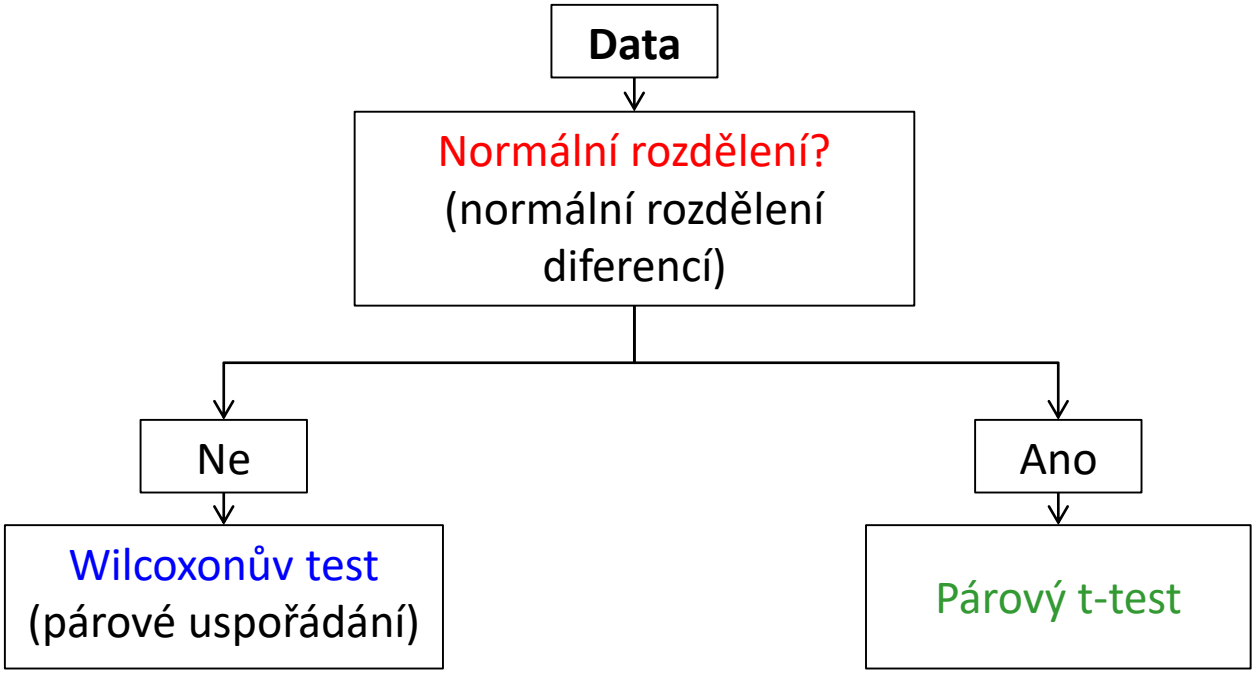
# Parametrické a neparametrické testy pro kvantitativní data – přehled

| Typ srovnání | Parametrický test | Neparametrický test |
|---|---|---|
| **1 skupina dat s referenční hodnotou – jednovýběrové testy:** | Jednovýběrový t-test, jednovýběrový z-test | Wilcoxonův test |
| **2 skupiny dat párově – párové testy:** | Párový t-test | Wilcoxonův test, znaménkový test |
| **2 skupiny dat nepárově – dvouvýběrové testy:** | Dvouvýběrový t-test | Mannův-Whitneyův test, mediánový test |
| **Více skupin nepárově:** | ANOVA | Kruskalův- Wallisův test |

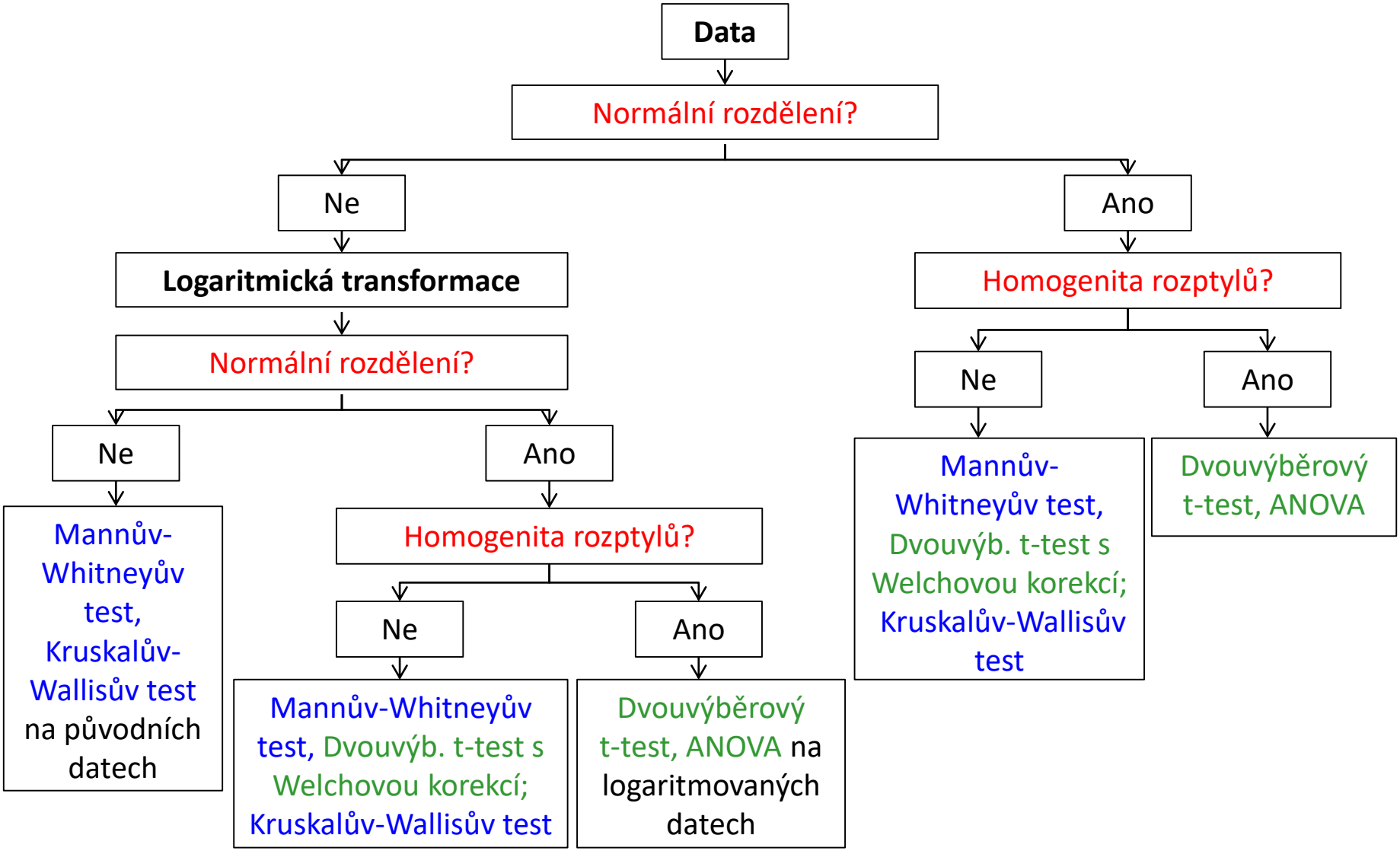
# Schéma při testování pomocí jednovýběrových testů



# Schéma při testování pomocí párových testů



# Schéma při testování dvou a více skupin



# Kontingenční tabulka

- Frekvenční sumarizace dvou binárních, nominálních nebo ordinálních proměnných.
- Obecně: **R x C kontingenční tabulka** (R – počet kategorií jedné proměnné, C – počet kategorií druhé proměnné).
- Speciální případ: 2 × 2 tabulka = čtyřpolní tabulka.

- Př.: Sumarizace vyšetřených osob podle typu onemocnění a věkových kategorií.

| Typ onemocnění | Věk | | | | Celkem |
|---|---|---|---|---|---|
| | <60 let | 60-70 let | 70-80 let | ≥80 let | |
| CN | 1 | 7 | 176 | 46 | 230 |
| MCI | 13 | 85 | 201 | 107 | 406 |
| AD | 9 | 34 | 90 | 64 | 197 |
| Celkem | 23 | 126 | 467 | 217 | 833 |

# Kontingenční tabulky – absolutní četnosti, řádková, sloupcová a celková procenta

## Kontingenční tabulka absolutních četností

| Skupina | Věk | | | | Celkem |
|---|---|---|---|---|---|
| | <60 let | 60-70 let | 70-80 let | ≥80 let | |
| CN | 1 | 7 | 176 | 46 | 230 |
| MCI | 13 | 85 | 201 | 107 | 406 |
| AD | 9 | 34 | 90 | 64 | 197 |
| Celkem | 23 | 126 | 467 | 217 | 833 |

## Kontingenční tabulka řádkových procent

| Skupina | Věk | | | | Celkem |
|---|---|---|---|---|---|
| | <60 let | 60-70 let | 70-80 let | ≥80 let | |
| CN | 0,4 | 3,0 | 76,5 | 20,0 | 100,0 |
| MCI | 3,2 | 20,9 | 49,5 | 26,4 | 100,0 |
| AD | 4,6 | 17,3 | 45,7 | 32,5 | 100,0 |
| Celkem | 2,8 | 15,1 | 56,1 | 26,1 | 100,0 |

## Kontingenční tabulka sloupcových procent

| Skupina | Věk | | | | Celkem |
|---|---|---|---|---|---|
| | <60 let | 60-70 let | 70-80 let | ≥80 let | |
| CN | 4,3 | 5,6 | 37,7 | 21,2 | 27,6 |
| MCI | 56,5 | 67,5 | 43,0 | 49,3 | 48,7 |
| AD | 39,1 | 27,0 | 19,3 | 29,5 | 23,6 |
| Celkem | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 |

## Kontingenční tabulka celkových procent

| Skupina | Věk | | | | Celkem |
|---|---|---|---|---|---|
| | <60 let | 60-70 let | 70-80 let | ≥80 let | |
| CN | 0,1 | 0,8 | 21,1 | 5,5 | 27,6 |
| MCI | 1,6 | 10,2 | 24,1 | 12,8 | 48,7 |
| AD | 1,1 | 4,1 | 10,8 | 7,7 | 23,6 |
| Celkem | 2,8 | 15,1 | 56,1 | 26,1 | 100,0 |

# Kontingenční tabulky – hypotézy

- Kontingenční tabulky umožňují testování různých hypotéz:
- **Nezávislost a shoda struktury** (Pearsonův chí-kvadrát test, Fisherův exaktní test)
  - Jeden výběr, dvě charakteristiky nebo více výběrů, jedna charakteristika – obdoba nepárového uspořádání
  - Př.: pacienti s AD – pohlaví × vzdělání (VŠ, SŠ, ZŠ); pacienti s AD v několika nemocnicích × věková struktura
- **Symetrie** (McNemarův test)
  - Jeden výběr, opakovaně jedna charakteristika – obdoba párového uspořádání
  - Př.: MMSE v normě a pod normou na začátku studie a dva roky po zahájení studie

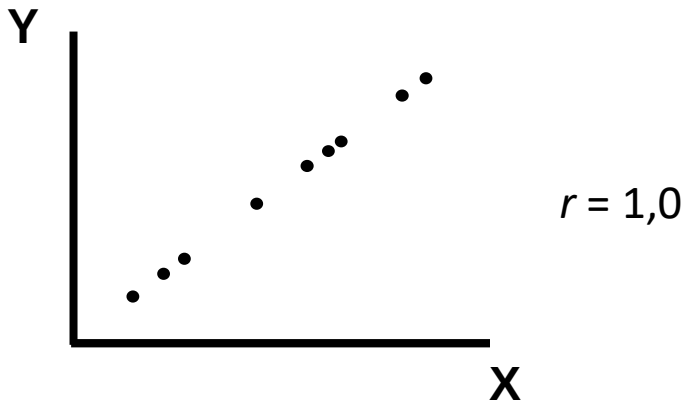
# Korelace

- **Korelační koeficient** – kvantifikuje míru vztahu mezi dvěma spojitými proměnnými (X a Y).

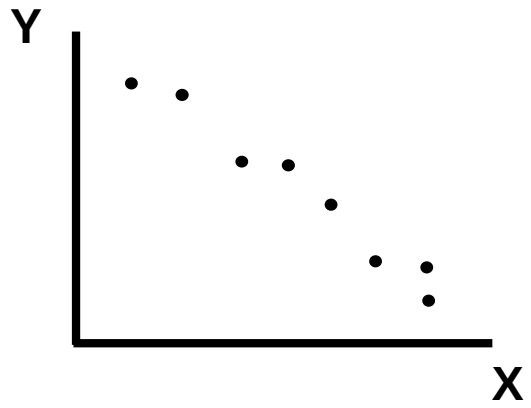
- Standardní metodou je výpočet **Pearsonova korelačního koeficientu (r)**:
  - Charakterizuje **linearitu** vztahu mezi X a Y – jinak řečeno variabilitu kolem lineárního trendu.
  - Nabývá hodnot od -1 do 1.
  - Hodnota r je kladná (kladná korelace), když vyšší hodnoty X souvisí s vyššími hodnotami Y, a naopak je záporná (záporná korelace), když nižší hodnoty X souvisí s vyššími hodnotami Y.
  - Proměnné jsou nekorelované, pokud *r* = 0.
  - Hodnoty 1 nebo -1 získáme, když body x-y grafu leží na přímce.

- Lze statistickým testem **otestovat, zda jsou dvě spojité proměnné nezávislé** – hypotézy mají tvar: $H_0: r = 0$ (tzn. korelační koeficient je roven nule) a $H_1: r \neq 0$.
- Neparametrická obdoba: **Spearmanův korelační koeficient** – není náchylný k odlehlým hodnotám (pracuje s pořadími pozorovaných hodnot).

# Pearsonův korelační koeficient (*r*)

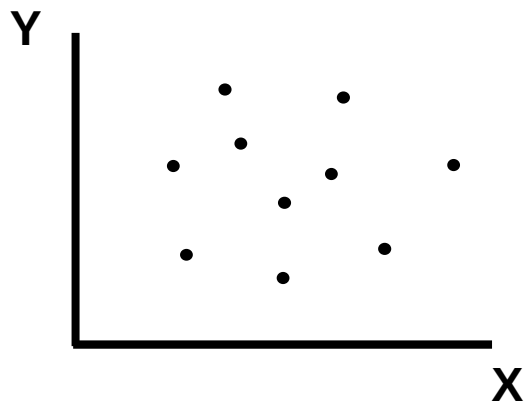



*r* = 1,0




*r* = -0,9


*r* = 0,4

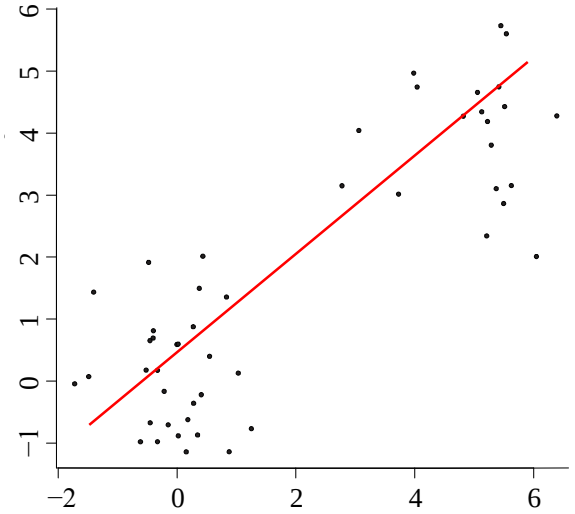



*r* = 0,05

# Pearsonův korelační koef. – problematické situace I.

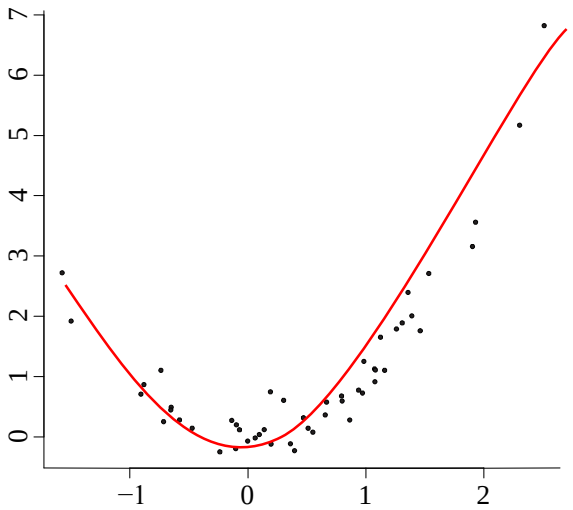
- Pearsonův korelační koeficient není vhodné počítat v situaci, kdy:
  - se v datech vyskytuje více skupin
  - proměnné mají nelineární vztah
  - se v datech vyskytují odlehlé hodnoty

**Více skupin**



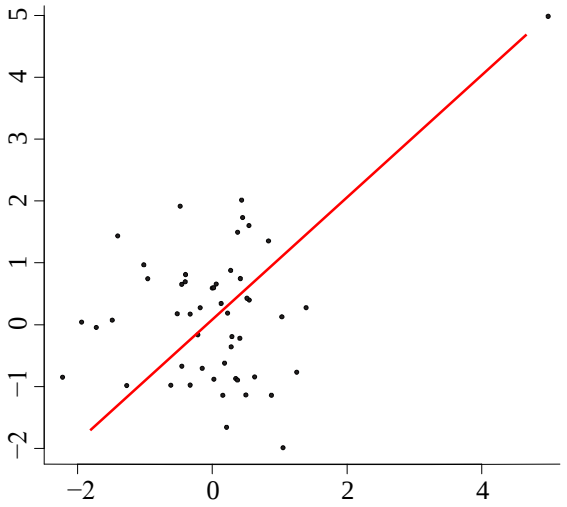
$r = 0{,}84$
($p < 0{,}001$)

**Nelineární vztah**



$r = 0{,}58$
($p < 0{,}001$)

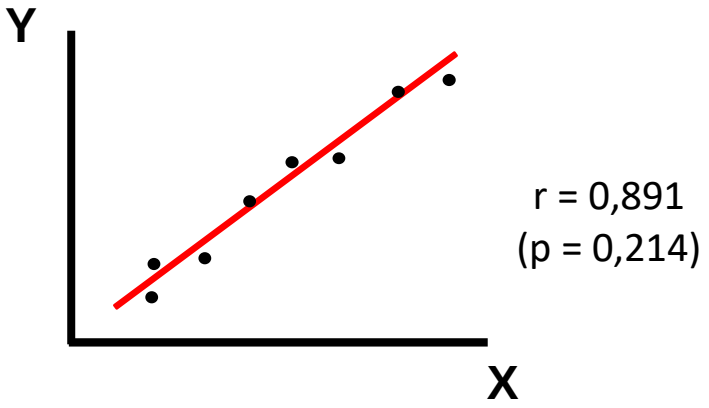
**Odlehlá hodnota**



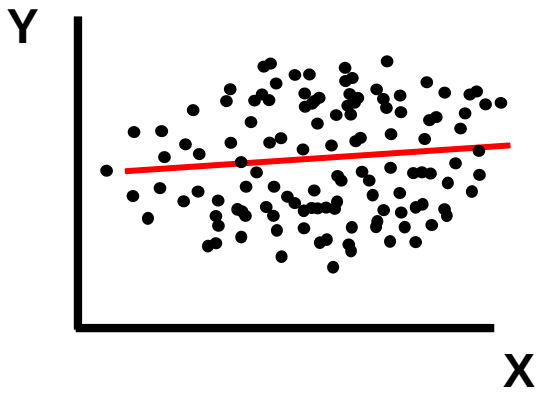
$r = 0{,}36$
($p = 0{,}009$)

# Pearsonův korelační koef. – problematické situace II.

- Problém velikosti vzorku:




r = 0,891
(p = 0,214)

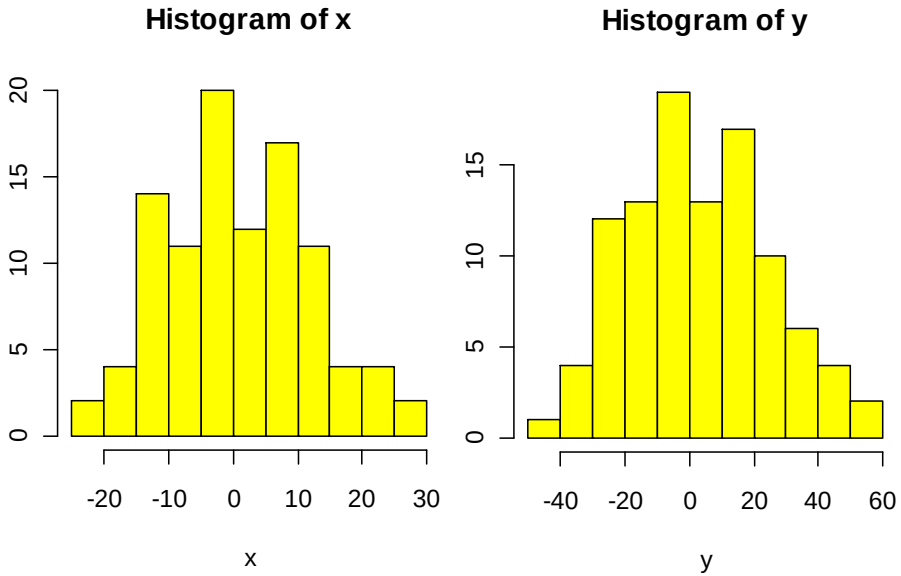



r = 0,012
(p = 0,008)

- Test na ověření, zda je Pearsonův korelační koeficient různý od nuly, je parametrický test – předpoklad normality srovnávaných spojitých proměnných!

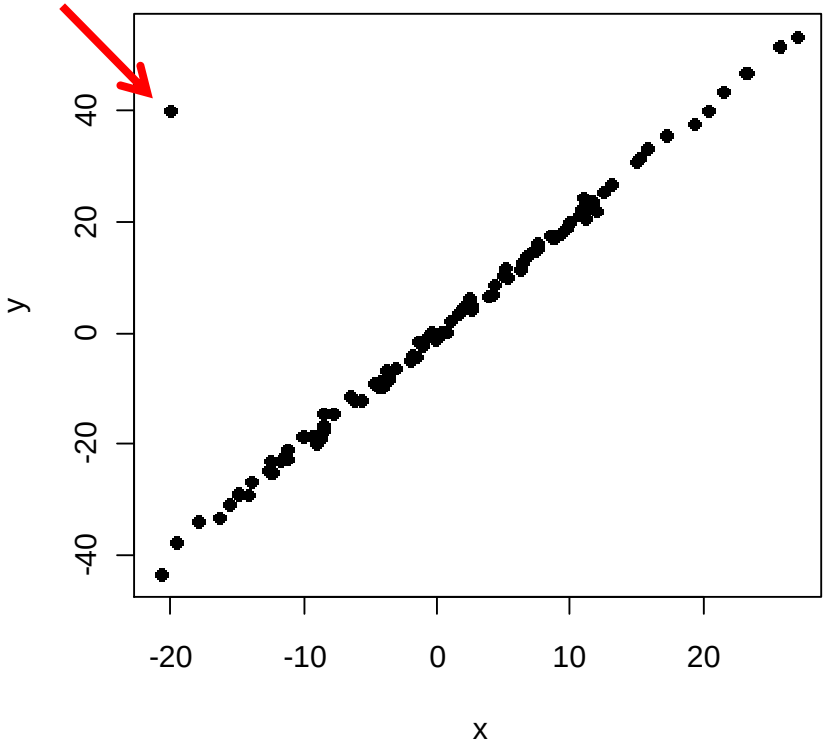
# Pearsonův korelační koef. – problematické situace III.

- Při srovnání dvou spojitých proměnných je nutné vykreslovat bodový graf, protože histogramy pro jednotlivé proměnné zvlášť nám nemusejí odhalit odlehlé hodnoty!





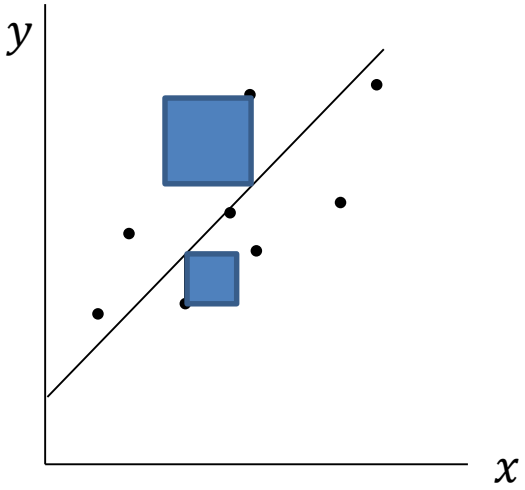




# Lineární regrese

Obecný zápis: $\boldsymbol{y} = \boldsymbol{X} * \boldsymbol{\beta} + \boldsymbol{\varepsilon}$

Zápis, pokud máme pouze jednu nezávisle proměnnou: $\boldsymbol{y} = \beta_0 + \beta_1 * \boldsymbol{x} + \boldsymbol{\varepsilon}$




**Odhad koeficientů β metodou nejmenších čtverců:**

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

***y*** – závisle proměnná (vysvětlovaná proměnná)

***x*** – nezávisle proměnná (vysvětlující proměnná, regresor)

***ε*** – náhodná složka modelu přímky (rezidua přímky)

$\beta_0$ – intercept
$\beta_1$ – regresní koeficient – „sklon regresní přímky“

# Shrnutí zásad při testování

1. **Znát základní typy testů** a vědět, pro jaká data se používají.
2. **Ověřit předpoklady testu** – smysl má pouze aplikace „správného“ testu na „správná“ data.
3. Posoudit, zda je výsledek **významný i z klinického hlediska**.
4. Být si vědom toho, že **statistický test není nic víc než matematický vzorec** aplikovaný na data, tedy existuje nenulová pravděpodobnost, že výsledek bude chybný (viz chyba I. a II. druhu). Ovlivnit výsledky testu můžeme například změnou velikosti vzorku.