Matematika pro geometrickou morfometrii (5)

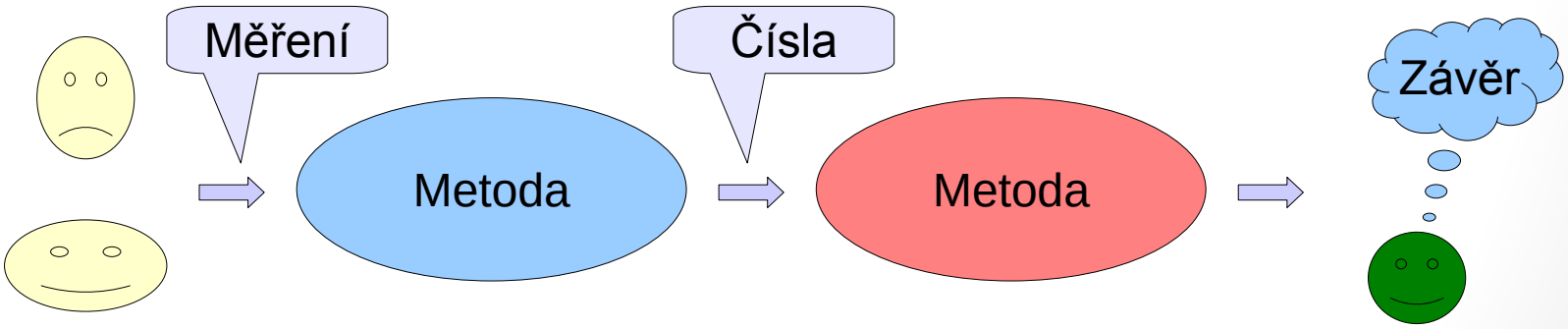
Ján Dupej (jdupej@cgg.mff.cuni.cz)

Laboratoř 3D zobrazovacích a analytických metod
Katedra antropologie a genetiky člověka
Přírodovědecká fakulta UK v Praze

Computer
Graphics
Charles
University

Cíle GM

1. Popsat tvar čísly
2. Čísla statisticky vyhodnotit



Doporučený software

- PAST – Paleontological Statistics
  - Tabulkový editor
  - Nabídka Statistics – základní testy
  - Nabídka Multivar – multivariační analýza
  - Nabídka Model – regresní analýza
- R, Matlab, Octave
  - Psaní skriptů
  - Cokoliv
- Excel
- Morphome3cs
  - „R in disguise“

Analýza dat

- Visualizace hrubých naměřených dat
  - Grafy
  - Scatter plot

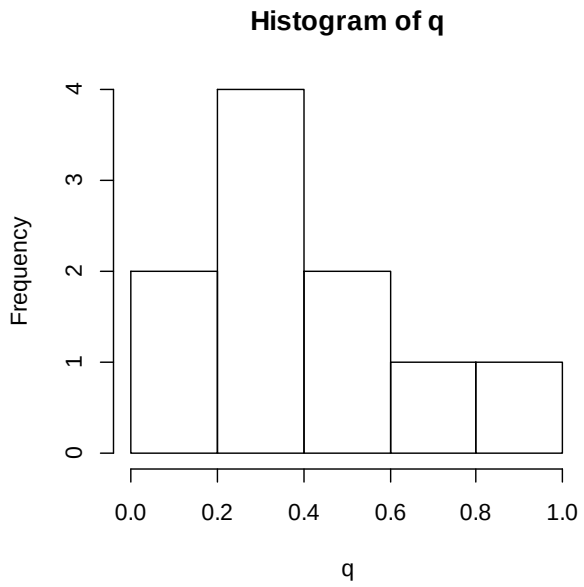
Analýza dat

- Souhrny
  - Průměr
  - Směrodatná odchylka
  - Median
  - Kvantily

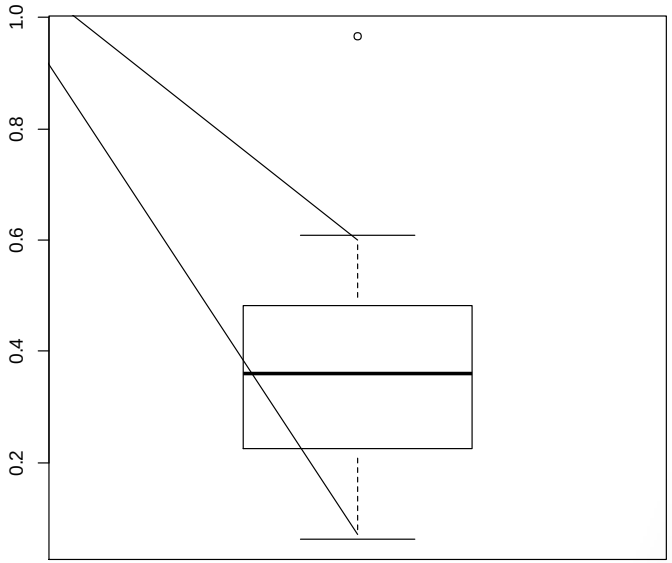
```
> q
 [1] 0.06288092 0.22550026 0.60736743 0.34576066 0.11118783 0.41442509
0.48277405 0.96585162 0.34863972 0.37133069
> mean(q)
[1] 0.3935718
> sd(q)
[1] 0.2590044
> summary(q)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.06288 0.25560 0.36000 0.39360 0.46570 0.96590
```

Analýza dat

- Histogram



- Boxplot s kníry



Pravděpodobnost

- Náhodný jev
  - Výsledek pro: hod mincí, hod kostkou...
  - Pravděpodobnost náhodného jevu
- Náhodná veličina
  - „Funkce na množině elementárních jevů“
  - Přirazení čísel jevům
- Rozdělení
  - Popis náhodné veličiny
  - Odpovídá histogramu pro mnoho opakování

Náhodná veličina

- Spojitá nebo diskrétní
- Popis tzv. **momenty**

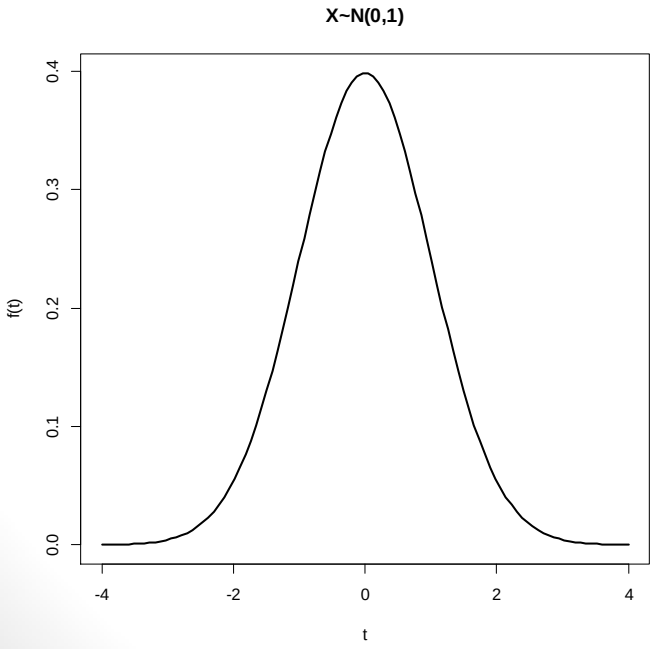
- Střední hodnota $EX = \int_{\infty}^{-\infty} x \cdot f(x)dx$
  - Aritmetický průměr $\bar{x} = \frac{1}{N}\sum x_i$

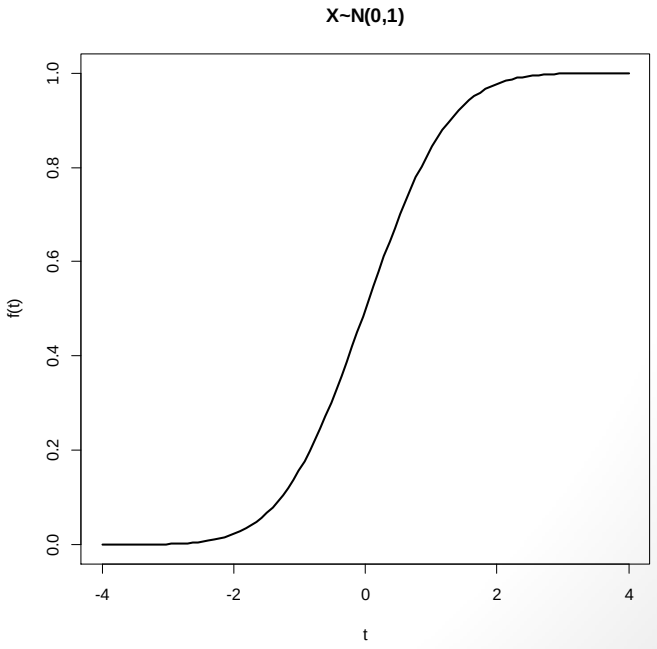
- Rozptyl $var(X) = E\left((X - EX)^2\right)$
  - Výběrový rozptyl $\sigma^2 = \frac{1}{N-1}\sum(x_i - \bar{x})^2$

- Šikmost (skewness)
- Špičatost (curtosis)

Hustota vs. Distribuční funkce

- Hustota
  - $f(t) \coloneqq P(X = t)$
  - Pst. že náhodná veličina X má hodnotu t
- Distr. funkce
  - $f(t) \coloneqq P(X \leq t)$
  - Pst. že náhodná veličina X má hodnotu nanejvýš t





Normální rozdělení

- Velké odchylky od očekávání jsou málo pravděpodobné

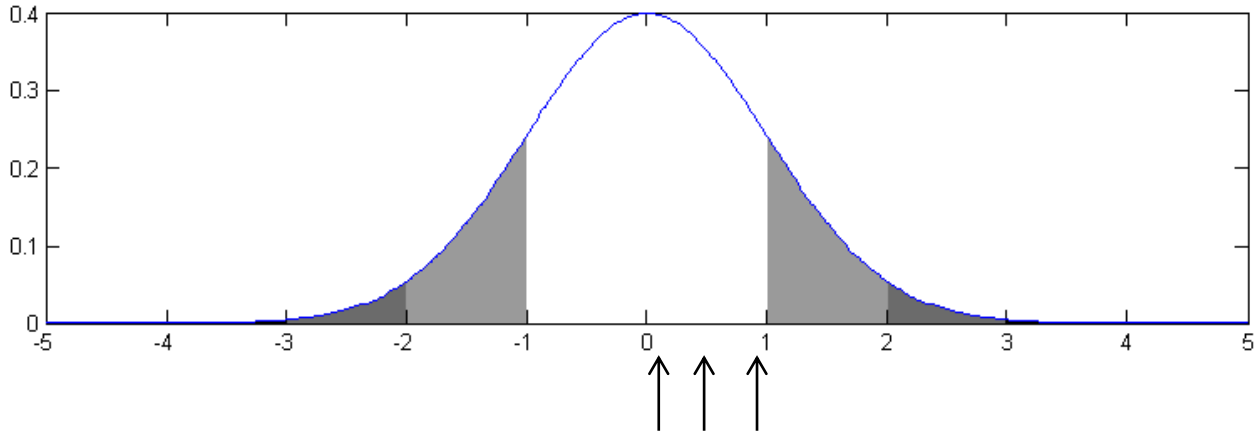
$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-X_0)^2}{2\sigma^2}}$$

- Centrální limitní věta
  - Komplexní děje se skládají z mnoha náhodných událostí – normální rozdělení je všude
  - Součty libovolného rozdělení se blíží normálnímu rozdělení

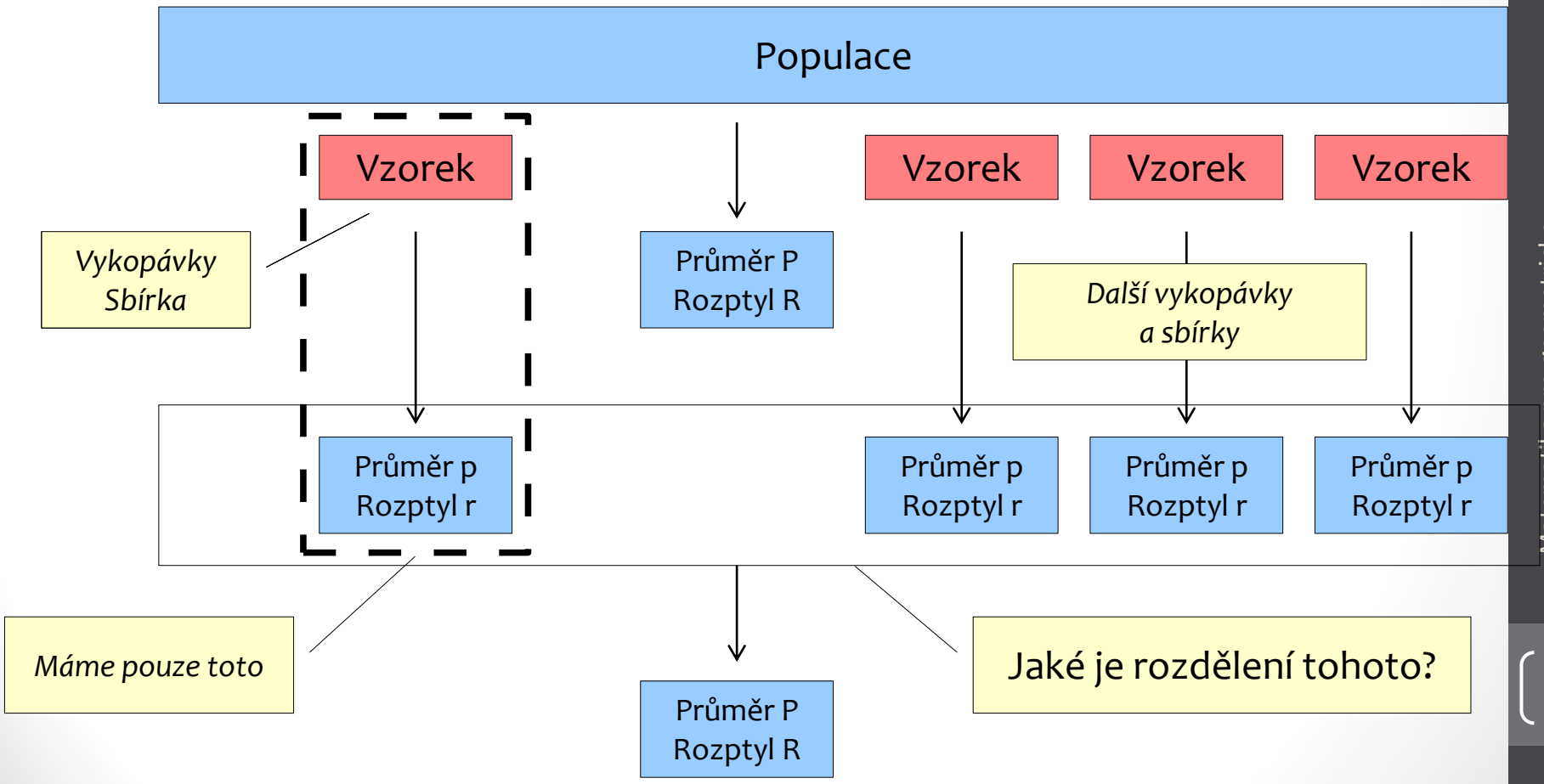
- Z-transformace
  - Posunutí a zúžení rozdělení aby výsledá střední hodnota byla 0 a kvadratická chyba 1
  - $Z = \frac{X-\mu}{\sigma(X)}$

Normální rozdělení

- Z-skóre
  - Symetrie rozdělení, tabulky
  - Celková ploch pod grafem = 1



Populace a vzorek



Populace a vzorek

- Rozdělení pravděpodobnosti výběru
  - Normální rozložení průměru pro velký vzorek, bez ohledu na rozdělení populace
  - Střední hodnota odpovídá střední hodnotě populace
  - Rozptyl nepřímo úměrný velikosti vzorku

- Větší vzorek = menší chyba odhadu

t - rozdělení

- Rozdělení pravděpodobnosti rozptylu při výběru vzorku z populace
- Vzorek je malý a neznáme rozptyl populace
  - Použijeme výběrovou směrodatnou odchylku vzorku
- Parametrizované velikostí vzorku
  - Vhodné pro GM kde se často pracuje s malými vzorky
- Komplikovaný výpočet distr. funkce
  - Tabulka
- t-hodnota

Intervalové odhady

- Chceme určit výsledek s předem danou přesností
  - Ptáme se na výsledek v celé populaci – vzorek
  - Skutočný výsledek se nalézá okolo průměru vzorku
  - Rozptyl odhadu je dán rozptylem populace a velikostí vzorku
  - Rozptyl v populaci nahradíme rozptylem ve vzorku
  - K výpočtu intervalu použijeme tabulkové hodnoty
    - Pro velké vzorky normální rozdělení
    - Pro malé vzorky t-rozdělení

Test hypotézy

- Vyvrátit pravdivost nějakého tvrzení o datech
  - Rovnost středních hodnot dvou vzorků
  - Rovnost střední hodnoty konkrétní hodnotě

- Jediný důkaz pro vyvrácení je v datech
- **Nevyvrátit ≠ potvrdit**

- Postup
  - Nulová hypotéza: X má střední hodnotu 0
  - Alternativní hypotéza: X nemá střední hodnotu 0
  - Určení skóre jevu popírajícího nulovou hypotézu
  - Výpočet p-hodnoty, **pravděpodobnosti že pozorovaný jev je dílem náhody**
  - Porovnání s hladinou významnosti (nejčastěji 0.05, 0.1)

Příklad – jednovýberový t-test

- Data
  - $(-4.4\ -0.3\ 7.6\ 2.7\ 4.4\ -0.2\ 1.5\ 0.3\ 3.0\ 0.0\ 0.2)$
- Ho: Data mají střední hodnotu 0
- Málo vzorků – použij t-rozdělení a t-statistiku
  - $t = \frac{\bar{X}-\mu_0}{s/\sqrt{N}} = \frac{1.345-0}{0.931} = 1.445$
- Porovnání t-statistiky s kritickou hodnotou
  - $n-1$ stupňů volnosti, kvantil dle hladiny významnosti
  - (Ne)Odmítnutí Ho
  - $t_{0.975}(10) = 2.23$ – Ho neodmítáme

Statistická významnost

- p-value
  - Nejnižší hladina na které ještě hypotézu nezamítáme
  - Porovnání přímo s hladinou významnosti

- Hvězdičková konvence
  - R, Morphome3cs

```
Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

Síla testu

- Chyba prvního typu
  - Test odmítl pravdivou hypotézu
  - „False negative“
  - Míra chyby α
- Chyba druhého typu
  - Test neodmítl neplatící hypotézu
  - „False positive“
  - Míra chyby β
- Síla testu
  - $1 - \beta$

Typy testů

- Rovnost středních hodnot
  - t-test, Hotelling T2, Wilcoxonův test
- Stejný rozptyl
  - F-test
- Stejné rozdělení pravděpodobnosti
- Test normality
  - Shapiro-Wilk, Kolmogorov-Smirnoff
- Test outlierů
- ...

Porovnání dvou vzorků

- Nepárový t-test
- Dva vzorky z dvou populací – významný rozdíl?
- Vzorek každé populace je náhodná veličina
  - X_1, X_2
- Rozdíl středních hodnot je také náhodná veličina
  - $\mu_1 - \mu_2$
- Jaké je rozdělení, střední hodnota, rozptyl?
  - Pro velké vzorky normální – kritická hodnota
  - Kritická hodnota – výpočet intervalového odhadu stř. hodnoty

Porovnání dvou vzorků

- Směrodatná odchylka
  - Za předpokladu normality a nezávislosti
  - $s = \sqrt{\frac{\sigma^2(X_1)}{n_1} + \frac{\sigma^2(X_2)}{n_2}}$

- Porovnání jako test hypotézy
  - Vzorky mají stejnou střední hodnotu – rozdíl je nulový
  - $z = \frac{\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)}{s}$

Dvouvýběrový t-test

- Pro malé vzorky se používá statistika t-skóre
  - Předpokládá se normalita a nezávislost
  - Neznámý ale stejný rozptyl

- Pro odhad směrodatné odchylky kombinujeme rozptyly vzorků
  - $s = \sqrt{\frac{(n-1)\sigma^2(X_1)+(m-1)\sigma^2(X_2)}{n+m+2}\left(\frac{1}{n}+\frac{1}{m}\right)}$

- Porovnám s kvantilem t-rozdělení s $n + m - 2$ stupni volnosti
  - $t = \frac{(\bar{X}_1-\bar{X}_2)}{s}$

Neparametrické testy

- Co když data nejsou normálně rozdělená
  - Příklad: málo dat, patologie,...
  - Použití bootstrap simulace
  - Wilcoxonův test
- Co když vzorky nemají stejný rozptyl
  - Složený odhad pro směrodatnou odchylku nemá normální ani t-rozdělení
  - Použití bootstrap simulace pro určení hladiny významnosti
- Monte Carlo, Jackknife

Permutační test

- Co když je rozdíl středních hodnot malý a rozptyl příliš velký
  - t-test nemusí zamítnout hypotézu jen díky náhodě výběru
  - Hledáme test s lepší rozlišovací schopností

- Počet případů kdy rozdíl středních hodnot B a A ku celkovému počtu opakování je pravděpodobnost že rozdíl B a A je pří Ho (rovnost A a B) náhoda.

- $A = (-1;\ -10;\ 11; 2; 7; -2; -14; -2; 8; 3; 12; 0)$
  - $\bar{A} = 1.1667;\ \sigma^2 = 7.8605$
- $B = (-3;\ -2; 10; 4; -2; 1; 9; -2; 0; -1; 13; -7)$
  - $\bar{B} = 1.6667;\ \sigma^2 = 6.0653$

```
-1 -10 11 2 7 -2 -14 -2 8 3 12 0 -3 -2 10 4 -2 1 9 -2 0 -1 13 -7
```

```
2 -7 1 -2 -14 -2 0 -3 -1 11 -2 8 -2 -10 12 0 -2 9 10 7 4 -1 13 3
```

Permutace,
Opakovat n-krát

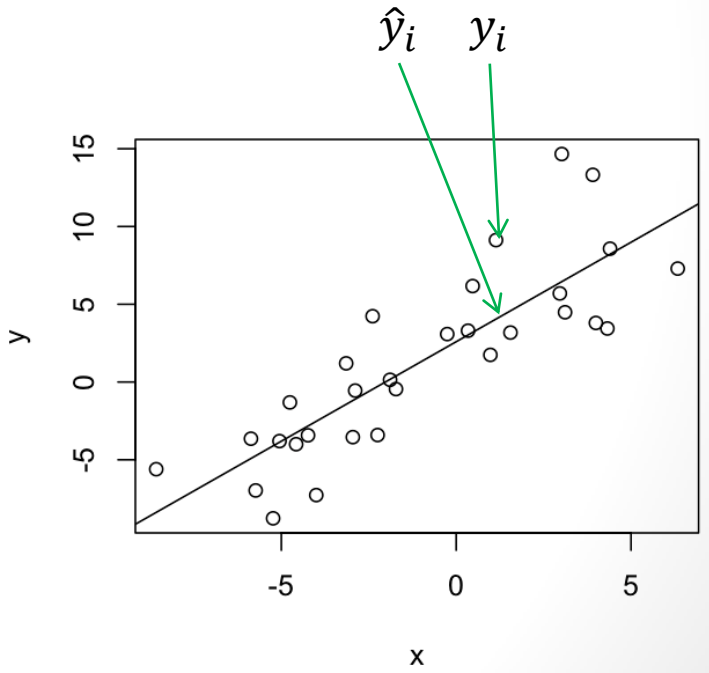
Návrh experimentu

- Párový test
  - Další možností je provést párový test/měření
  - Jedinec se vyskytuje v obou populacích
- Randomizace
  - Vzorky jsou vybrány **náhodně** při dělení do skupin
- Lokální kontrola
  - Eliminace variability
  - Rozdělení testované množiny do bloků
- Replikace
  - Stejně velké skupiny náhodných jedinců, stejný postup **opakovaného měření** a stejné výchozí podmínky

Regresní anaýza

- Vztah závislosti mezi dvěma veličinami
  - Závislá x a nezávislá y
- Jaký může být – tzv. model
  - Lineární
  - Vyšší stupeň
- Myšlenka – minimalizovat čtverec vzdálenosti vzorků v ose y

$$y = \alpha + \beta x$$

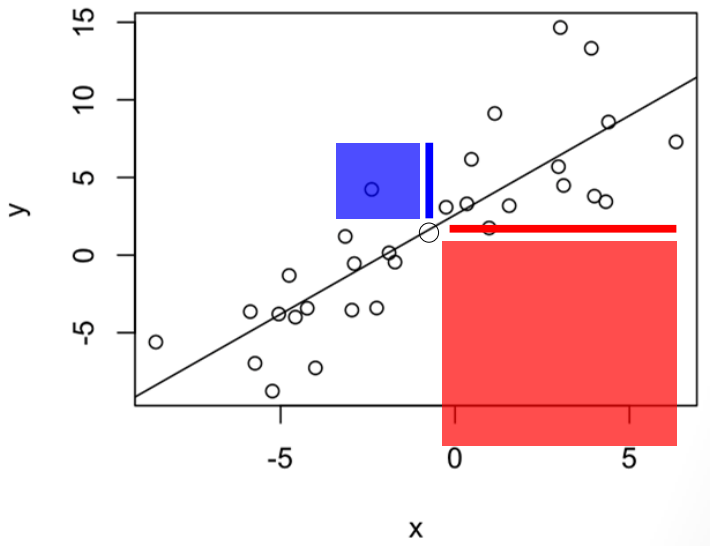


$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Regresní analýza

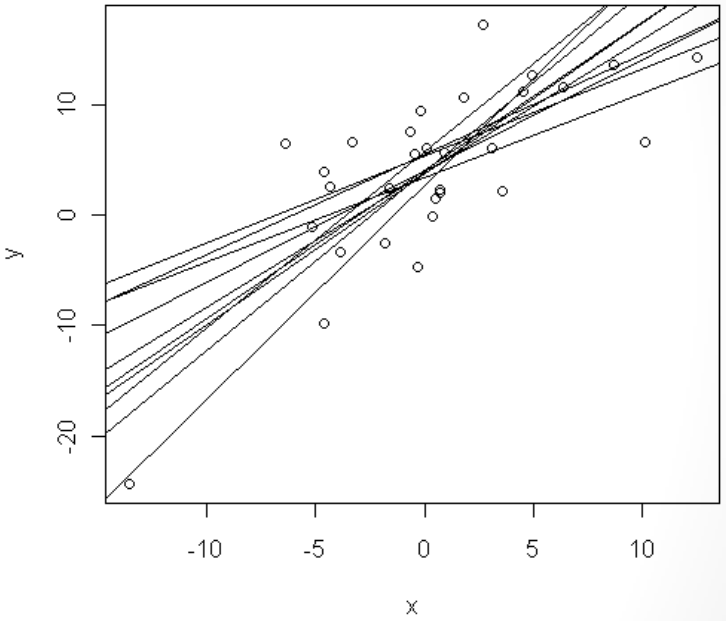
- Koeficienty
  - $\beta = \frac{SSxy}{SSxx}$
  - $\alpha = \bar{y} - \beta\bar{x}$

- Další míry
  - $SSxx = \sum(x_i - \bar{x})^2$
  - $SSyy = \sum(y_i - \bar{y})^2$
  - $SSyy = \sum(x_i - \bar{x})\,(y_i - \bar{y})$



Regresní analýza

- Čtverec korelace – **koeficient determinace**
- **Korelační koeficient**
  - Rostoucí/klesající regresní křivka
- Inferenční regresní analýza
  - Parametry křivky jsou náhodné veličiny
- Vícerozměrná regresní analýza
  - Více závislých proměnných, jedna nezávislá



Regresní analýa - R

```
> data()
> trees
   Girth Height Volume
1    8.3     70   10.3
2    8.6     65   10.3
3    8.8     63   10.2
4   10.5     72   16.4
5   10.7     81   18.8
6   10.8     83   19.7
7   11.0     66   15.6
8   11.0     75   18.2
9   11.1     80   22.6
10  11.2     75   19.9
11  11.3     79   24.2
...

> a<-trees$Girth
> b<-trees$Height
> lm(a~b)

Call:
lm(formula = a ~ b)

Coefficients:
(Intercept)            b
    -6.1884       0.2557
```

```
> summary(lm(a~b))

Call:
lm(formula = a ~ b)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2386 -1.9205 -0.0714  2.7450  4.5384

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.18839    5.96020  -1.038  0.30772
b            0.25575    0.07816   3.272  0.00276 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1

Residual standard error: 2.728 on 29 degrees of
freedom
Multiple R-squared:  0.2697,    Adjusted R-squared:
0.2445
F-statistic: 10.71 on 1 and 29 DF,  p-value: 0.002758
```

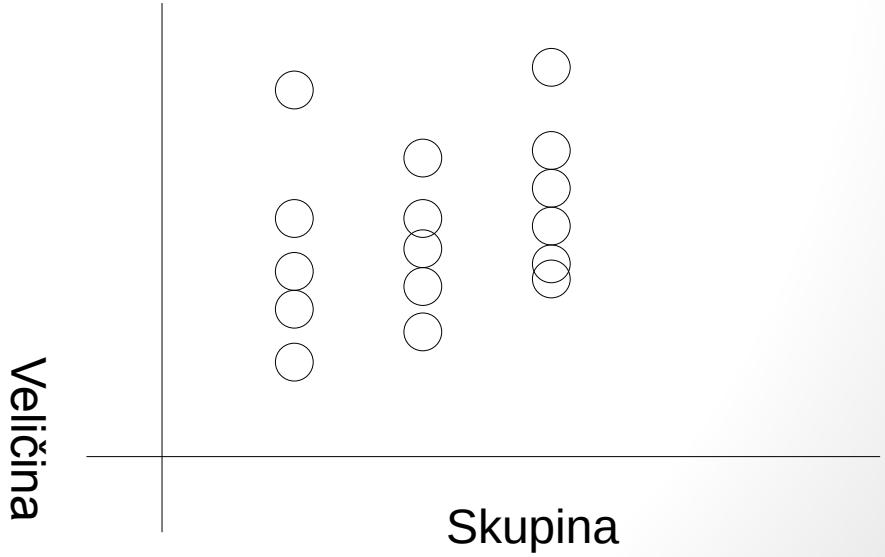
ANOVA

- Analysis of variance
  - Závislost veličiny popisující jedince na ostatních
  - Zobecňuje t-test pro více skupin (nezávislé proměnné jsou diskrétní, popisují kategorii)
  - Předpoklad normality a stejného rozptylu

- F-hodnota
  - $F = \frac{MS_b}{MS_e}$

- F-rozdělení

- Obtížné na výpočet



ANOVA - výpočet

- $MS_b = \frac{SS_b}{K-1} = \frac{\sum_i^K n_i(\bar{X}_i - \bar{X})^2}{K-1}$
- $MS_e = \frac{SS_e}{N-K} = \frac{\sum_i^K \sum_j^{n_i} (X_{i,j} - \bar{X}_i)^2}{N-K}$

Střední hodnota vzorku (Grand mean)

Střední hodnota skupiny

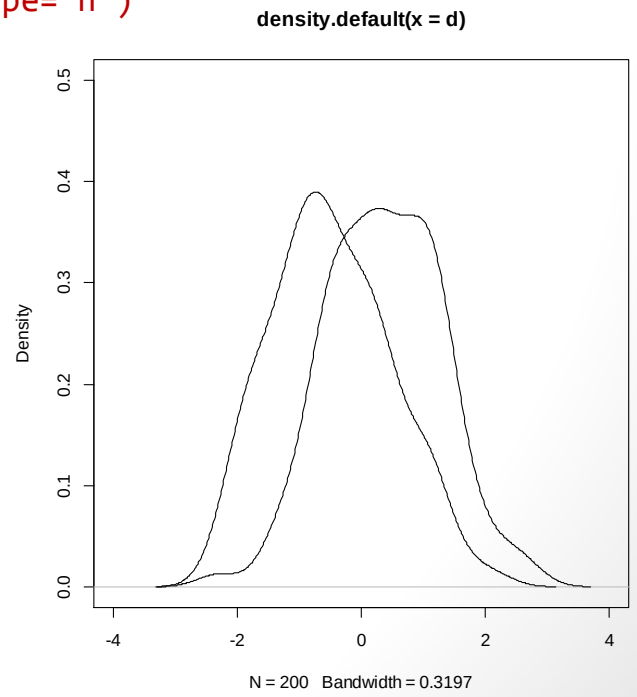
- P-hodnota se spočítá s využitím tabulky F-rozdělení
  - Existuje statistický rozdíl mezi skupinami?

ANOVA – interpretace

- ANOVA odhalí že existuje statisticky významný rozdíl, neřekne která skupina od které a jak moc
- Provádí se dodatečné testy každého s každým, podobně jako t-test
- Takových testů existuje víc, např. HSD-Tukey
  - Předpoklad stejného počtu prvků v každé skupině n_g
  - $Tukey_{HSD} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{MS_e / n_g}}$
- Koeficient se vyhodnotí např. pomocí tabulky
  - Podobně jako t, F-hodnota

ANOVA - R

```
> d <- c(rnorm(100, mean=-0.5, sd=1), rnorm(100, mean=0.5, sd=1))
> cls <- c(rep(0,100), rep(1,100))
> summary(aov(d ~ cls))
             Df Sum Sq Mean Sq F value   Pr(>F)
cls           1  35.97   35.97   41.13 1.02e-09 ***
Residuals   198 173.14    0.87
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> plot(density(d), xlim=c(-4,4), ylim=c(0,0.5), type="n")
> lines(density(d[1:100]))
> lines(density(d[101:200]))
```



Two-way ANOVA

- Pokud máme dvě kategorizující funkce (faktor)
  - Např. věkové skupiny a pohlaví
- Máme tři nulové hypotézy – získáme 3 p-hodnoty
  - Závislost veličiny na první kategorii, na druhé kategorii
  - Závislost první kategorie na druhé kategorii (interakce)
- Stejný počet jedinců v kombinaci kategorií
  - Jde i pokud není stejný počet, komplikovanější

MANOVA

- Zkoumáme případ více závislých proměnných na jedné nebo víc kategoriích

- Wilks lambda – ukáže že existuje významný vztah nezávislých závislých proměnných
  - Malá hodnota – dobrá separabilita

- Pro další zkoumání, např. diskriminační analýza

Hotellingův T^2-test

- Zobecnění t-testu pro multivariační analýzu (více proměnných)
- Je třeba z vektorových náhodných veličin získat skalární hodnotu (výsledek testu)

$$t^2 = n(\mathbf{x} - \mu)'\mathbf{W}^{-1}(\mathbf{x} - \mu)$$

Střední hodnota vzorku | Testovaná střední hodnota | Kovarianční matice vzorku

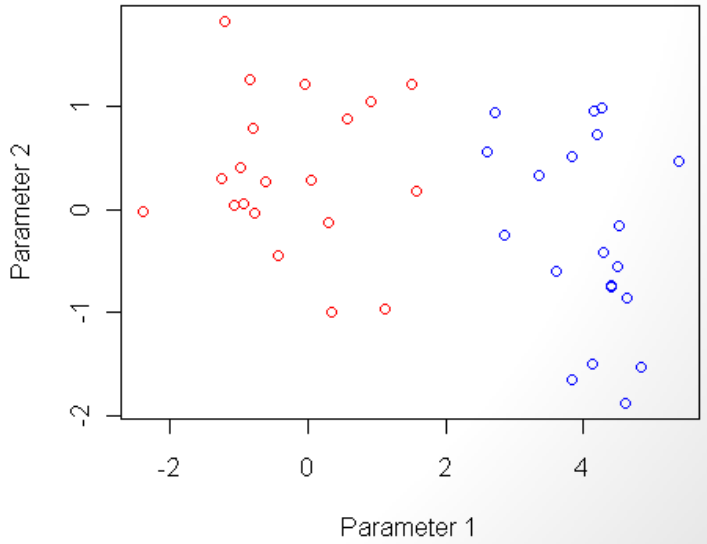
- t^2 má Hotellingovo t-kvadrát rozdělení (tabulka)
- Dvouvýběrová varianta

Diskriminační analýza

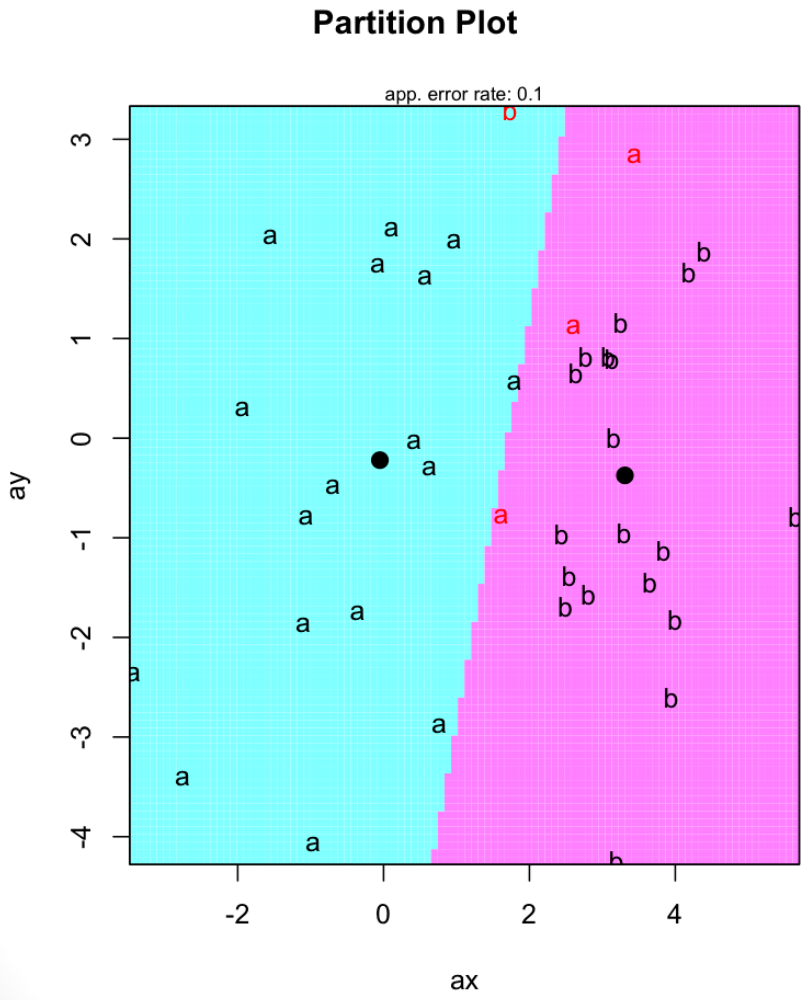
- Dichotomie – rozdělitelnost vzorku na dvě skupiny
- Hledání takové diskriminační funkce, která jedince x:
  - $f(x) > 0$ přiřadí do první skupiny
  - $f(x) < 0$ přiradí do druhé skupiny

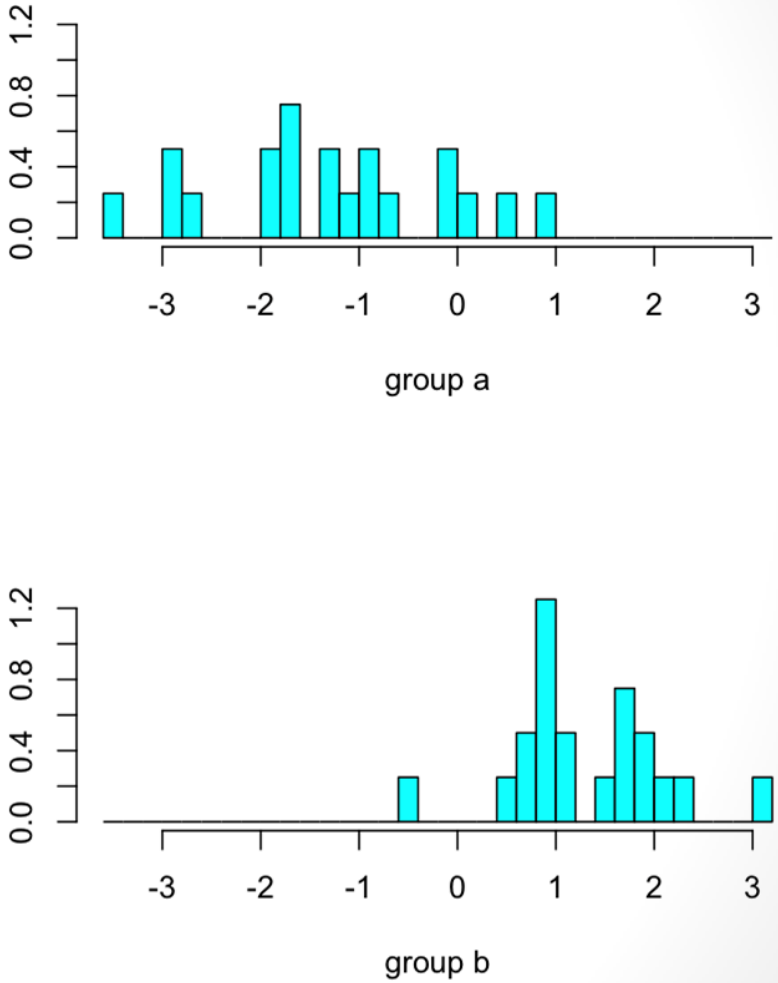
- Lineární diskriminační analýza (LDA) – f je lineární

- Dimenze dat je libovoln
- Rozšíření pro více skupin



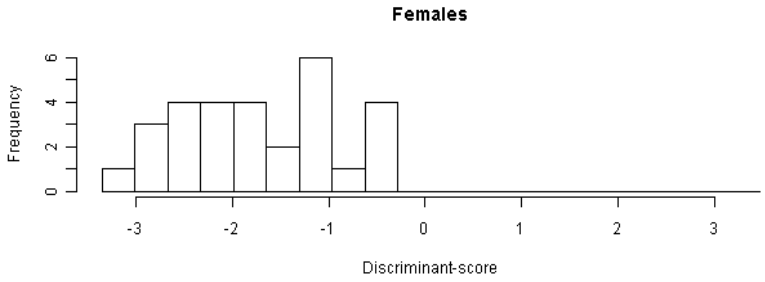
LDA

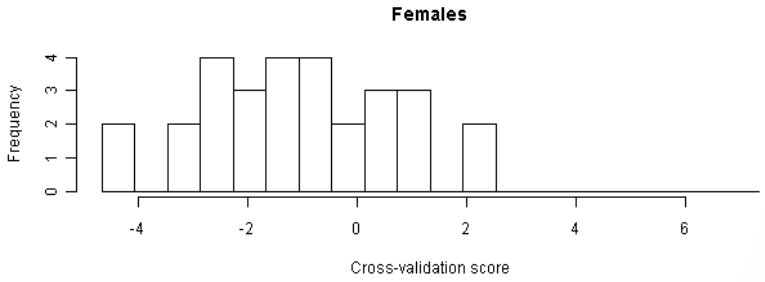


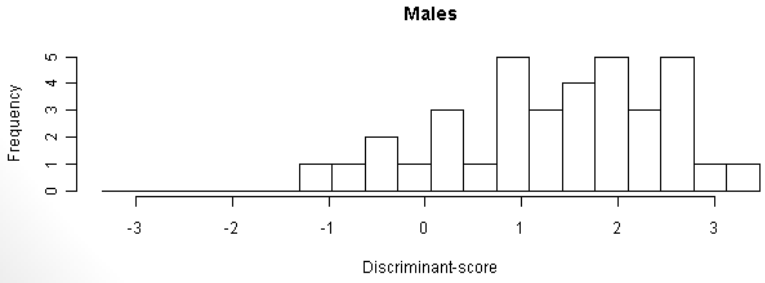


Interpretace DA

- Oddělení skupin ve vzorku je optimální pro vzorek, ne celou populaci – chceme výsledek zobecnit na populaci
  - **Cross-validation** – trénovací množina, testovací množina, počítá se úspěšnost na testovací množině
  - Varianty: k-fold, leave-one-out

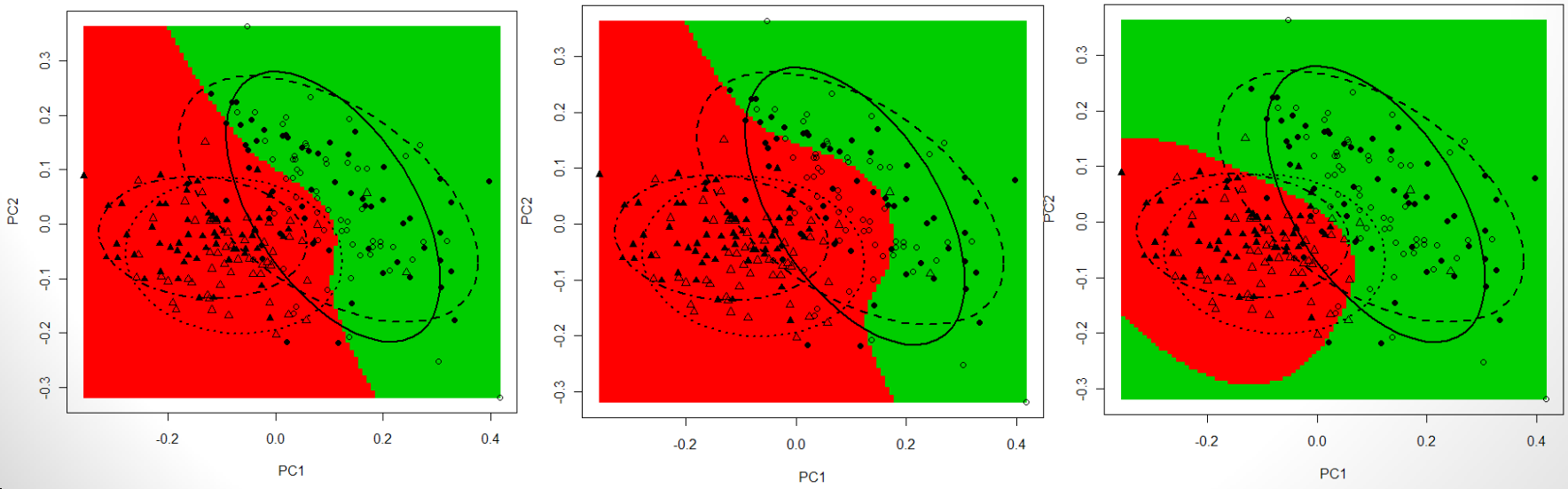






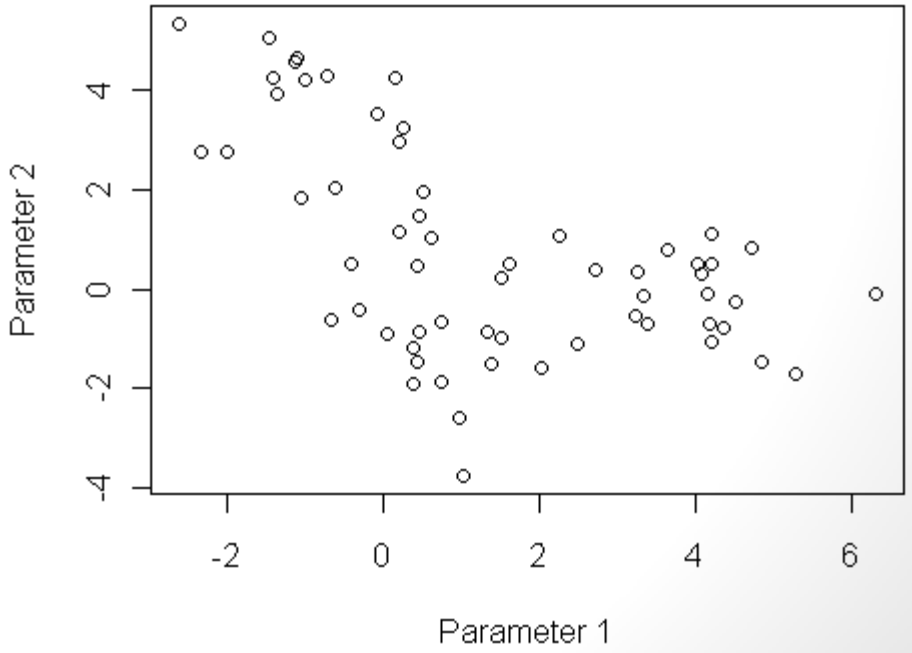
Modifikace DA

- Složitější podmínky
  - Support vector machines (SVM) – maximum margin criterion
- Složitější dělení prostoru
  - Quadratic discriminant analysis (QDA)
  - SVM kernels



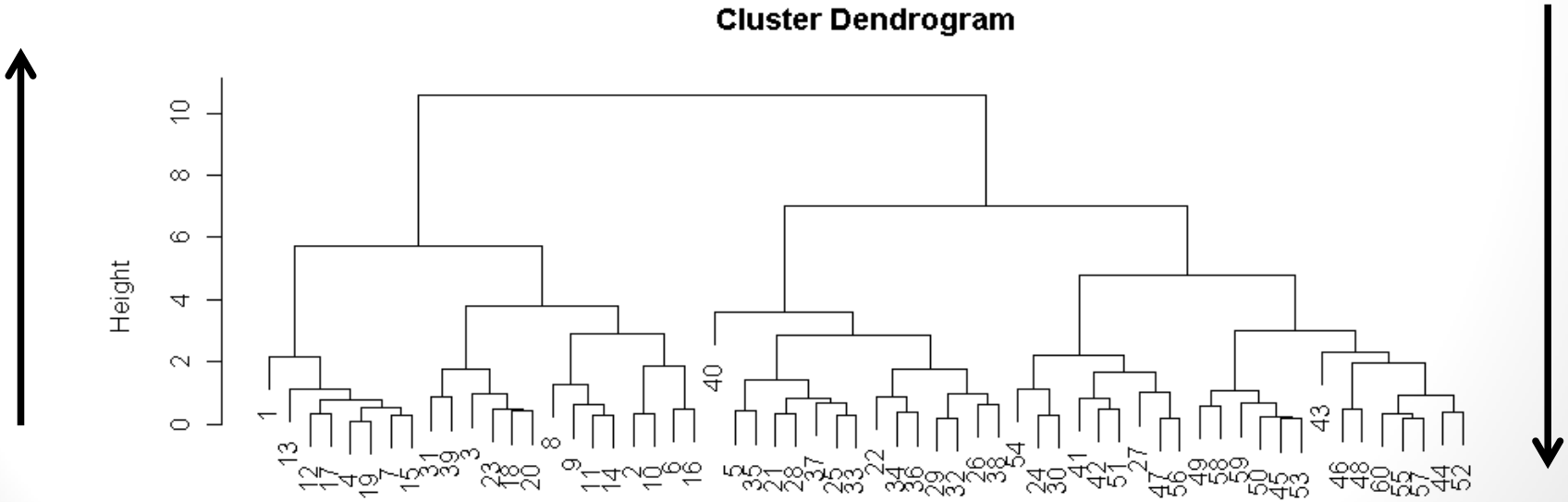
Shluková analýza

- Co když nevím
  - Kolik skupin data obsahují
  - Do kterých skupín data patří (učení bez učitele)
  - Jestli jsou shluky lineárně oddělitelné
- Hledání přirozených shluků (explorativní metoda)
- Libovolná dimenze
- Hierarchické shlukování
  - Míra podobnosti
- Nehierarchické shlukování
  - Znám počet shluků



Hierarchické shlukování

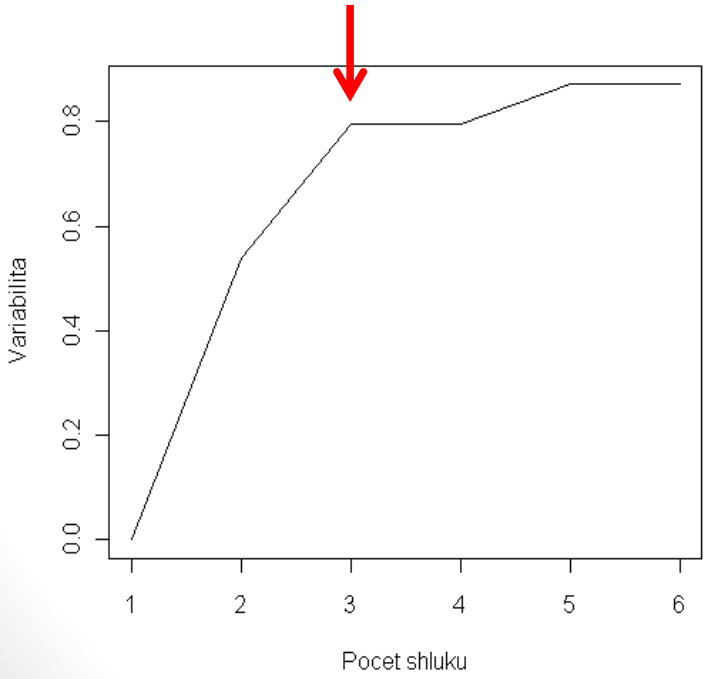
- Normalizace dat, použití vhodných metrik
- Aglomerativní, divizivní techniky

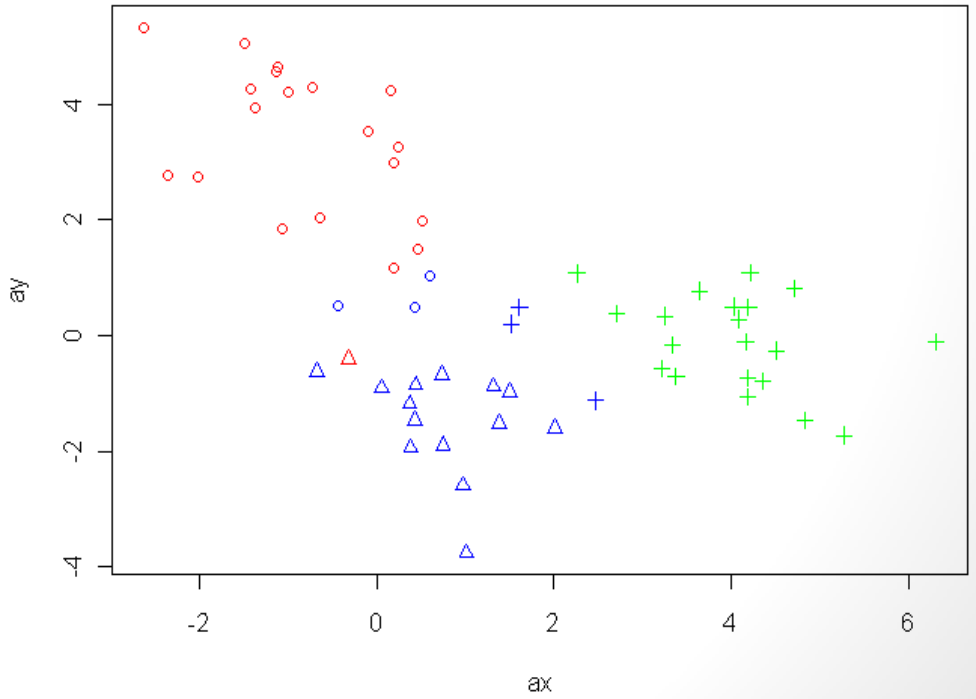


Hierarchické shlukování

- Vysvětlená variabilita je dána poměrem variability shluků k celkové variabilitě – **elbow criterion**

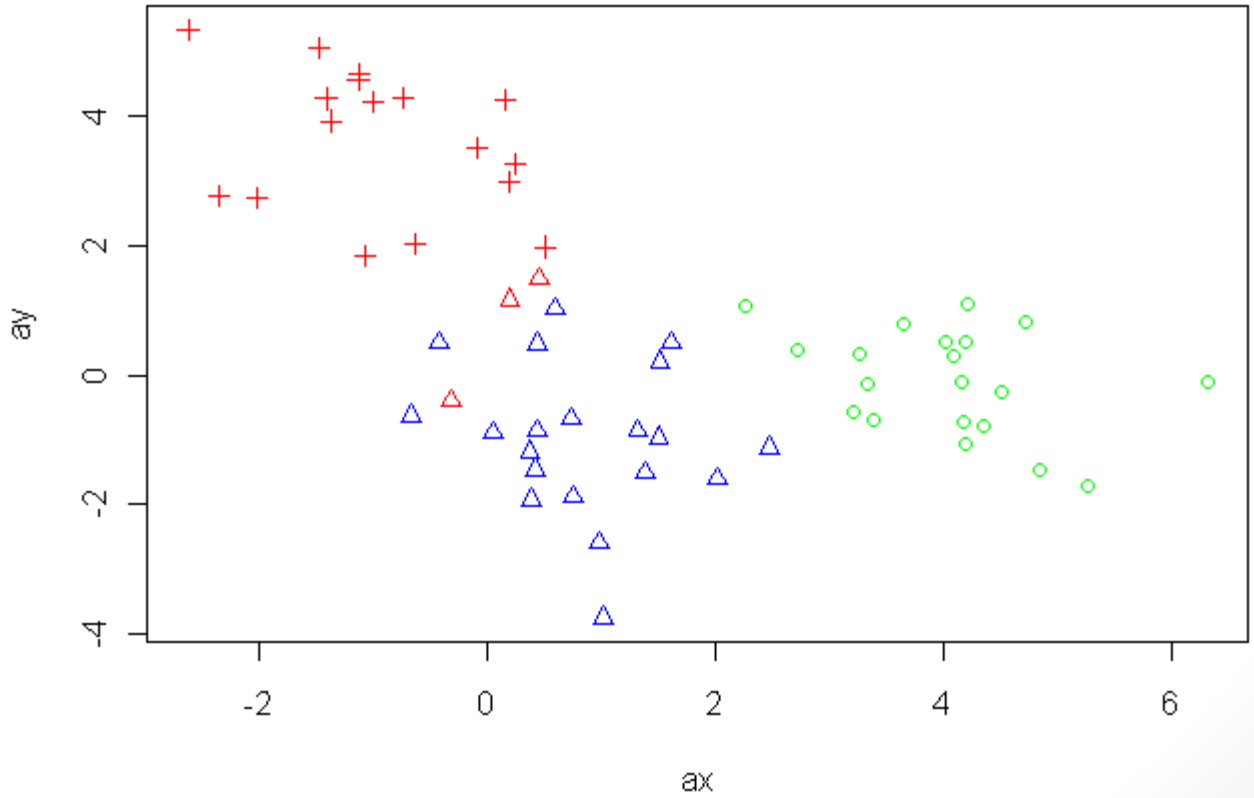
- Nalezení optimálního počtu shluků





K-means shlukování

- Znám K počet shluků
- Umístím (náhodně) **středy** a shlukuji nejbližší sousedy
- Přepočítám středy a opakuji



Klasifikace - interpretace

- Úspěšnost klasifikace
  - Počet správně zaklasifikovaných / počet celkem
- Posteriorní
  - Nacvičit na všech, zjistit počet správně klasifikovaných
- Cross-validace
  - Rozdělit data na „foldy“
  - Nacvičit na $k - 1$ a testovat zbytek, opakovat
  - Víc restriktivní, lépe vypovídá o vhodnosti klasifikátoru