Matematika pro geometrickou morfometrii

Václav Krajíček

Vaclav.Krajicek@mff.cuni.cz

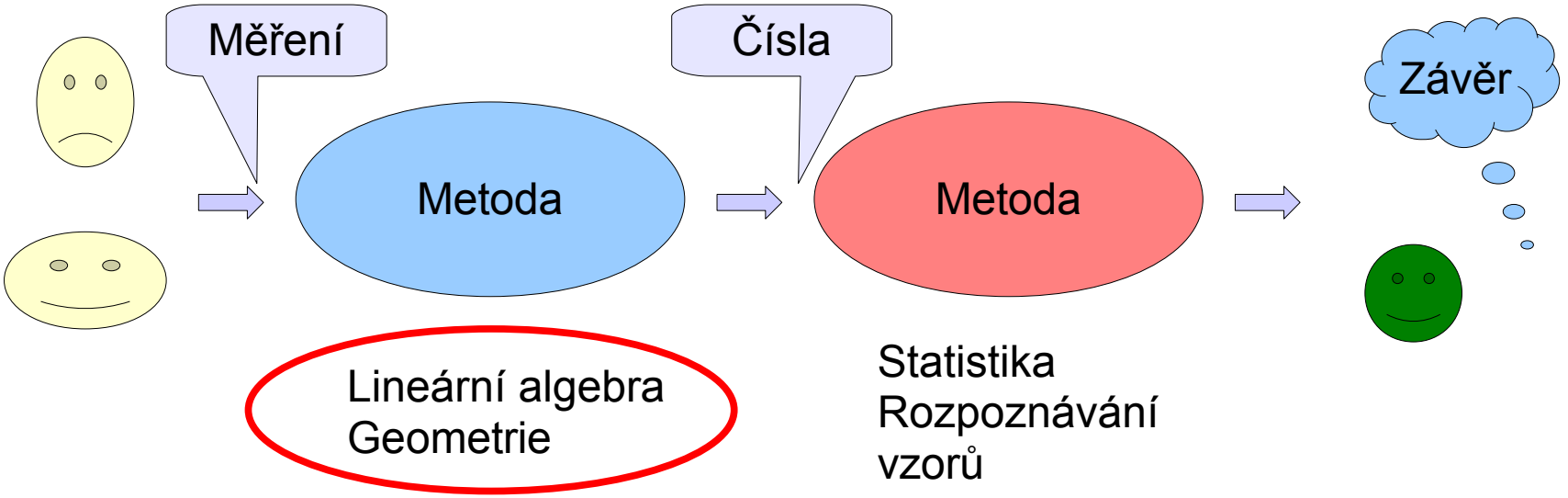
Department of Software and Computer Science Education
Faculty of Mathematics and Physics
Charles University

MFF UK
C G
G
Prague

Přednáška 4

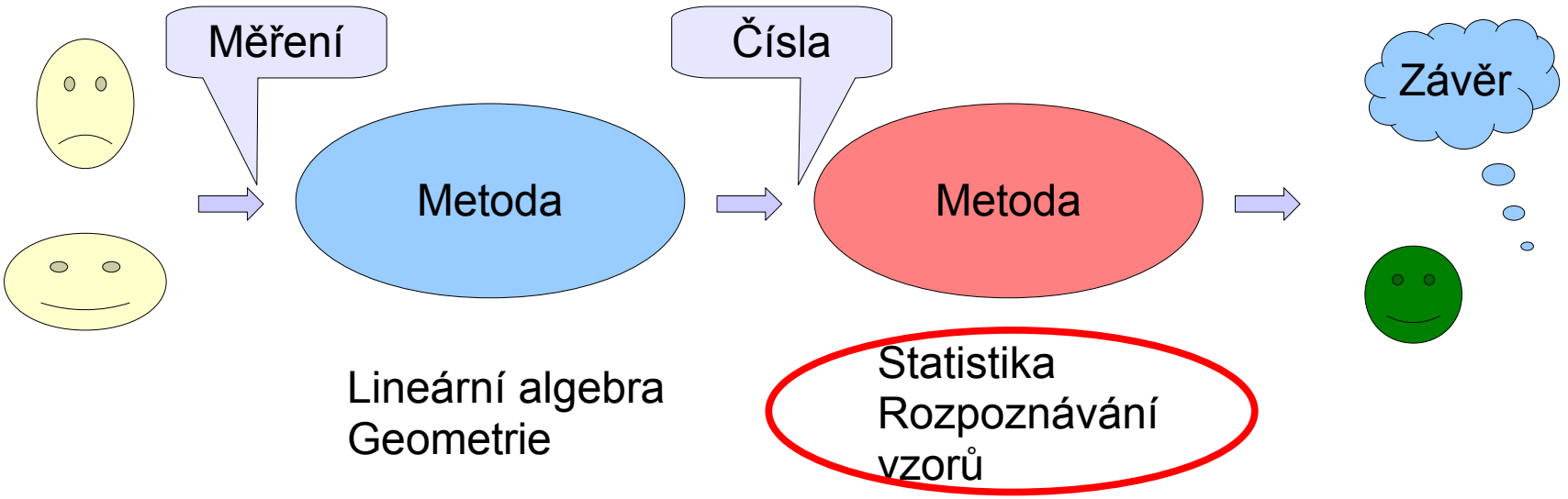
Opakování

- Naměřené souřadnice
- Určení tvarových proměnných (landmarkové metody)
  - Registrace (dvoubodová, GPA), Warps, Parametry statistického modelu (PCA)
- Zobrazení rozdílů (TPS, FESA)



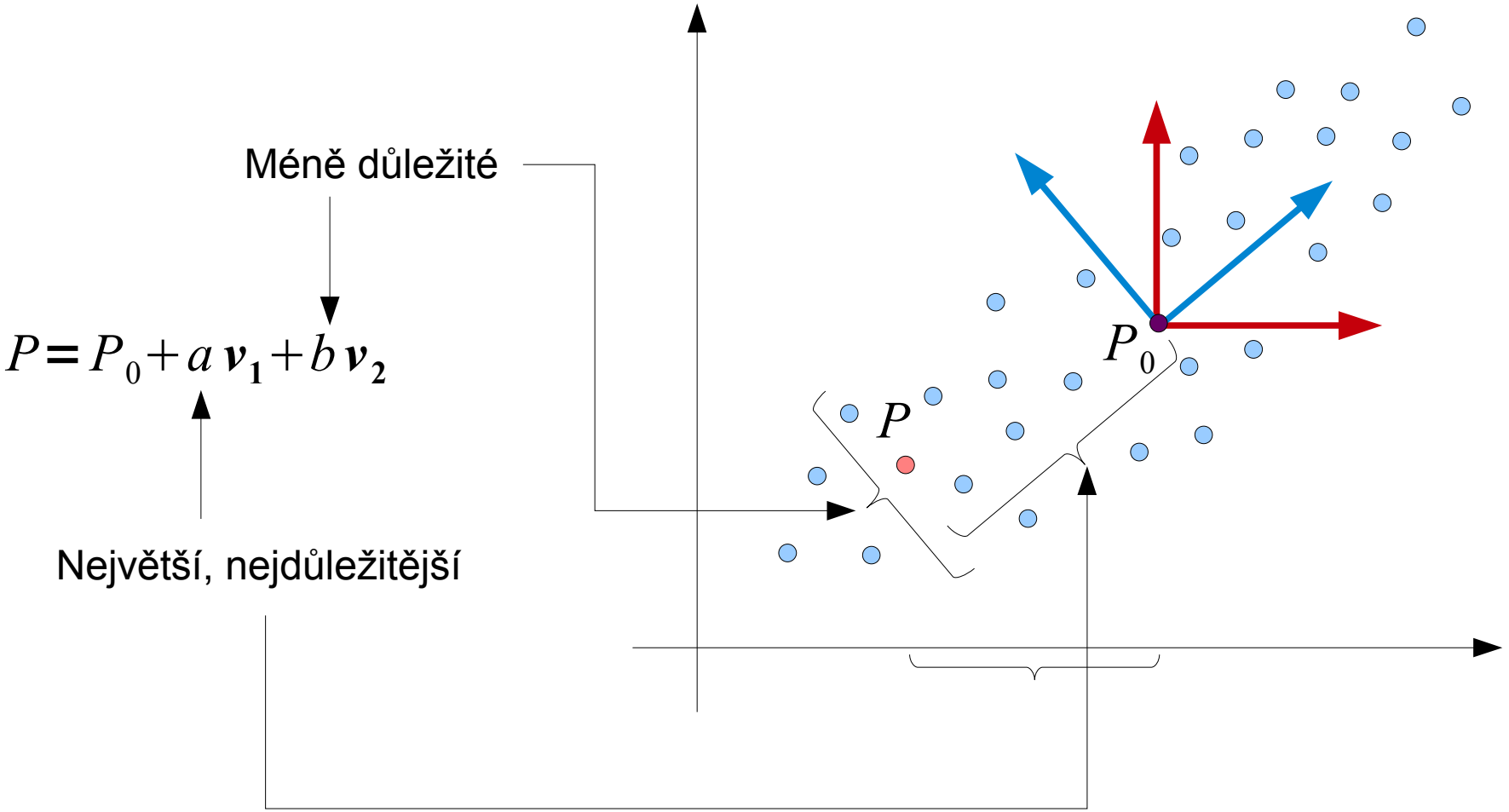
Opakování

- Statistické testy
- Klasifikace
- Skryté souvislosti pomocí regresní analýzy



Opakování - PCA

- Obvyklejší obrázek



PCA - výpočet

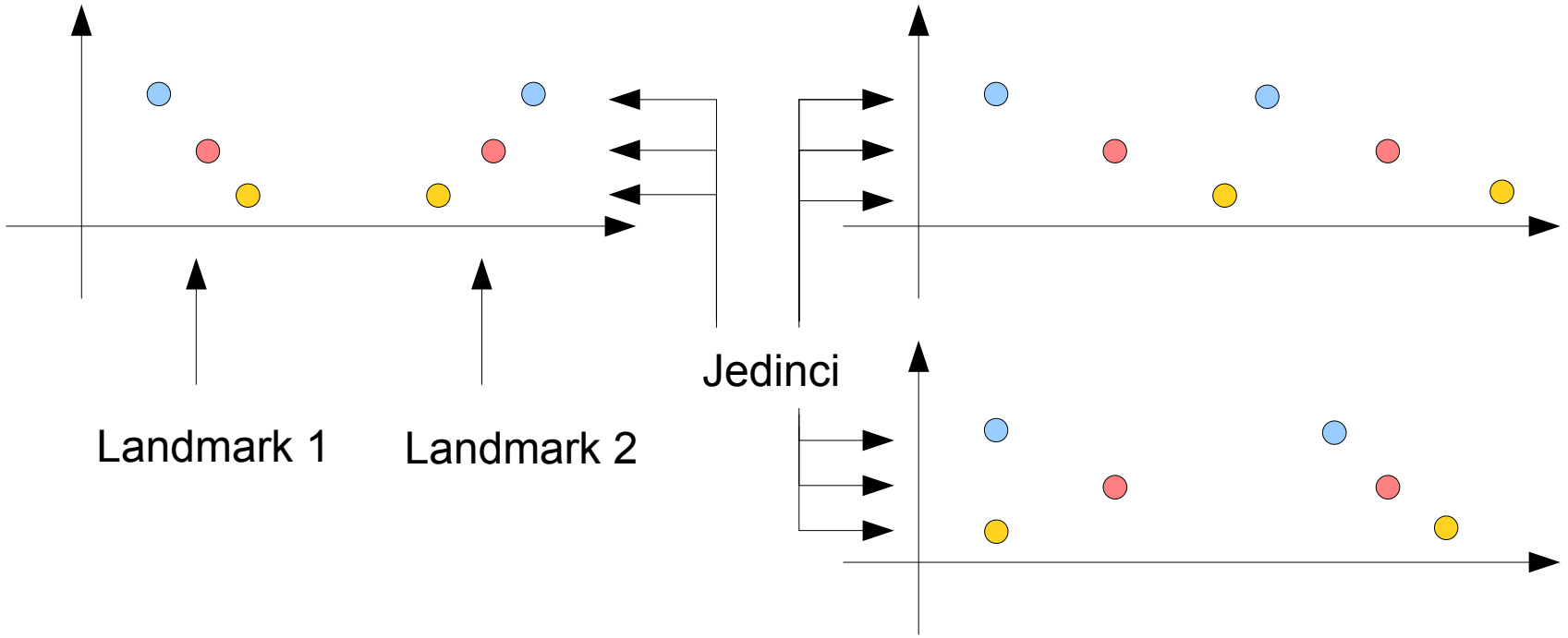
- Jak najít množinu hlavních komponent?
  - Rozsáhlá teorie, vynecháme
- Hlavní komponenty odpovídají vlastním vektorům kovarianční matice
- Kovariance popisuje „podobnost“ dvou landmarků
- Kovarianční matice „podobnost“ každého s každým

$$C_{ij} = \frac{1}{N} \sum_k (P_{ki} - \overline{P_i})(P_{kj} - \overline{P_j})$$

$$C = \begin{bmatrix} C_{00} & \cdots & C_{0n} \\ \vdots & \ddots & \vdots \\ C_{n0} & \cdots & C_{nn} \end{bmatrix}$$

Kovariance landmarků

- X-souřadnice landmarku 1 a 2



$$C_{12}=\frac{1}{3}((1-2)(10-9)+(2-2)(9-9)+(3-2)(8-9))=-2$$

$$C_{12}=\frac{1}{3}((1-3)(8-10)+(3-3)(10-10)+(5-3)(12-10))=8$$

$$C_{12}=\frac{1}{3}((1-2)(9-10)+(4-2)(10-10)+(1-2)(11-10))=0$$

PCA - výpočet

- Dvojic vlastní číslo a vektor je stejně jako jedinců
- Vlastní číslo určuje důležitost komponenty/vlastního vektoru

$$C\,\boldsymbol{x} = \lambda\,\boldsymbol{x}$$

- Můžu se rozhodnout kolik komponent do svého modelu chci zahrnout
  - Víc komponent → Víc (zbytečných) detailů, parametrů
  - Poměr součtu vlastních čísel vybraných komponent ku celkovému součtu odpovídá množství informace

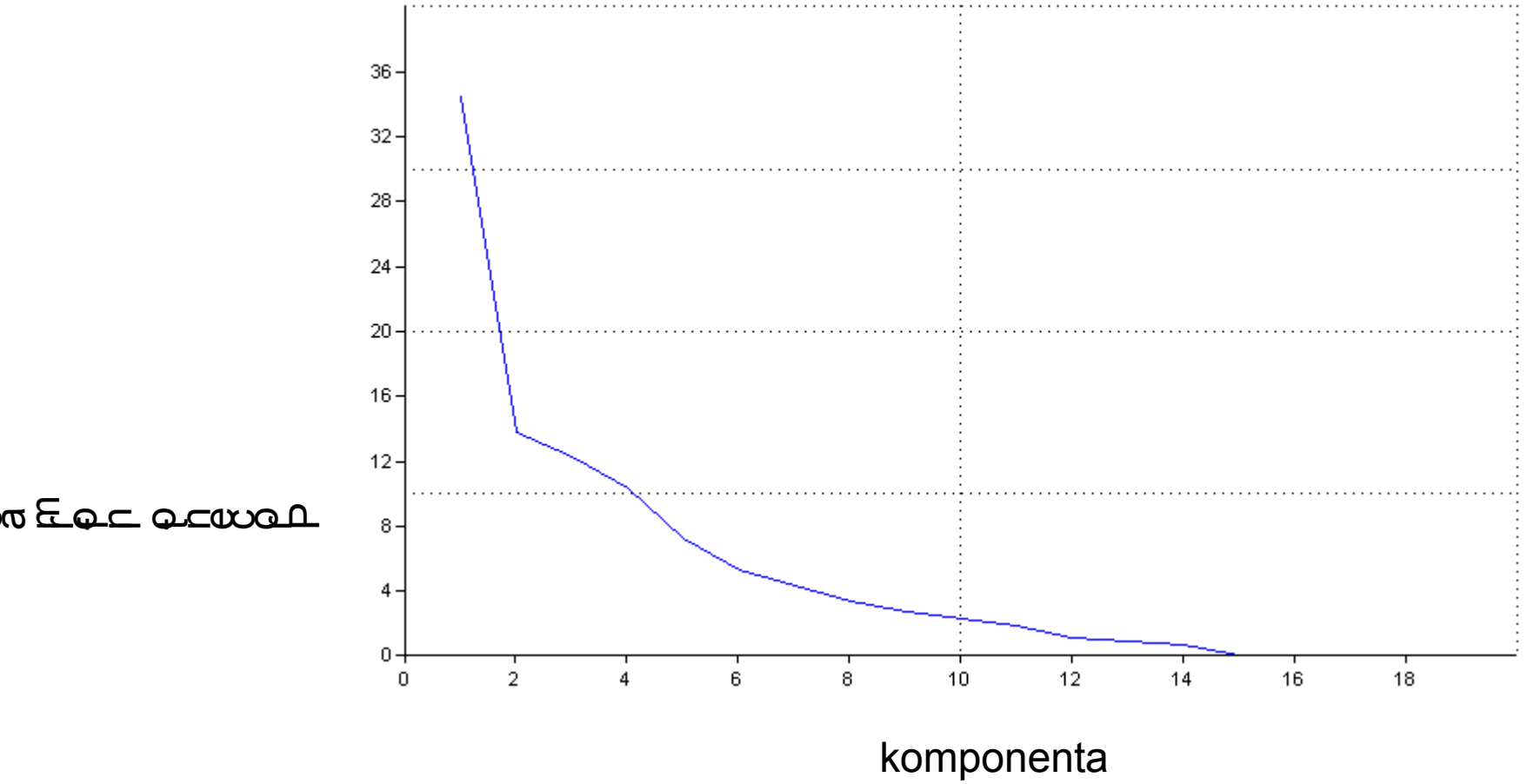
$\boldsymbol{v}_1 = [\ldots]$ $\quad \lambda_1 = 1 \quad \lambda_2 = 3.1 \quad \lambda_{total} = 11.45$

$\boldsymbol{v}_4 = [\ldots]$ $\quad \lambda_3 = 2 \quad \lambda_4 = 5.35 \quad \lambda_{4+2}/\lambda_{total} = 0.738$

73.8% informace

PCA – váha komponent

- Nejvíc informace se typicky nachází v prvních několika komponentách → vzorek není náhodný



PCA - příklad

- Data

$$P=\begin{bmatrix} 1 & 2 & 1.5 & 4 \\ 1.1 & 1.2 & 2.5 & 3.8 \\ 1.1 & 1.2 & 1.5 & 3.5 \\ 1.3 & 1.1 & 1.2 & 3.7 \end{bmatrix}$$

- Model

$$M = P_0 + \sum v_i k_i$$

- Vlastní vektory/čísla

$v_1=[-0.8117 \quad -0.3713 \quad -0.1295 \quad 0.4318]$	$\lambda_1=0$	0% informace
$v_2=[0.5388 \quad -0.2217 \quad -0.0343 \quad 0.8120]$	$\lambda_2=0.016$	2.87% informace
$v_3=[0.2167 \quad -0.8866 \quad -0.1194 \quad -0.3909]$	$\lambda_3=0.2114$	37.92% informace
$v_4=[0.0618 \quad 0.1642 \quad -0.9838 \quad -0.0377]$	$\lambda_4=0.3301$	59.21% informace

- Průměr $P_0=[1.1250 \quad 1.3750 \quad 1.6750 \quad 3.7500]$

PCA - příklad

- Zbývá dopočítat koeficienty (souřadnice, score) pro „namodelování“ původních dat

$$\boldsymbol{k_1} = \begin{bmatrix} 0.0000 & 0.0031 & -0.6580 & 0.2577 \end{bmatrix}$$

$$P_0 = \begin{bmatrix} 1.1250 & 1.3750 & 1.6750 & 3.7500 \end{bmatrix}$$

$$P_0 + \boldsymbol{v_4} k_{1,4} = \begin{bmatrix} 1.1409 & 1.4173 & 1.4215 & 3.7403 \end{bmatrix}$$

$$P_0 + \boldsymbol{v_3} k_{1,3} + \boldsymbol{v_4} k_{1,4} = \begin{bmatrix} 0.9984 & 2.0007 & 1.5001 & 3.9975 \end{bmatrix}$$

$$P_0 + \boldsymbol{v_2} k_{1,2} + \boldsymbol{v_3} k_{1,3} + \boldsymbol{v_4} k_{1,4} = \begin{bmatrix} 1.0000 & 2.0000 & 1.5000 & 4.0000 \end{bmatrix}$$

$$P_0 + \boldsymbol{v_1} k_{1,1} + \boldsymbol{v_2} k_{1,2} + \boldsymbol{v_3} k_{1,3} + \boldsymbol{v_4} k_{1,4} = \begin{bmatrix} 1.0000 & 2.0000 & 1.5000 & 4.0000 \end{bmatrix} = P_1$$

- Zobrazení vybraných 2 až 3 koeficientů do grafu pro všechny exempláře → scatter plot

PCA - demonstrace

- PCA na landmarky
  - Matlab/Octave
  - Past
- Které landmarky jsou nejvíce ovlivněné první komponentou?
  - Ty které mají v komponentě nejvyšší absolutní hodnoty přes všechny souřadnice

$$v_1 = [\underbrace{0.3742 \quad 0.0117}_{\text{první landmark}} \quad \underbrace{-0.4139 \quad -0.5919}_{\text{druhý landmark}} \quad \underbrace{0.0397 \quad 0.5802}_{\text{třetí landmark}}]$$

první landmark	druhý landmark	třetí landmark
0.3859	1.0058	0.6199

PCA - demonstrace

- PCA na partial warp scores
  - Matlab/Octave
    1) Víc jedinců, spočítám partial warps scores
    2) Průměrné score
    3) PCA na matici
  - tpsRelw

We should note that many of the studies applying PCA to geometric data call the method "relative warps analysis" (RWA). PCA and RWA are not exactly equivalent, because the components of variance extracted by RWA are sometimes weighted by bending energy (originally, RWA was an analysis of components of variation relative to bending energy, hence the term "relative" in the name of the method). When variation is not weighted by bending energy, RWA is PCA. We prefer the more familiar term.

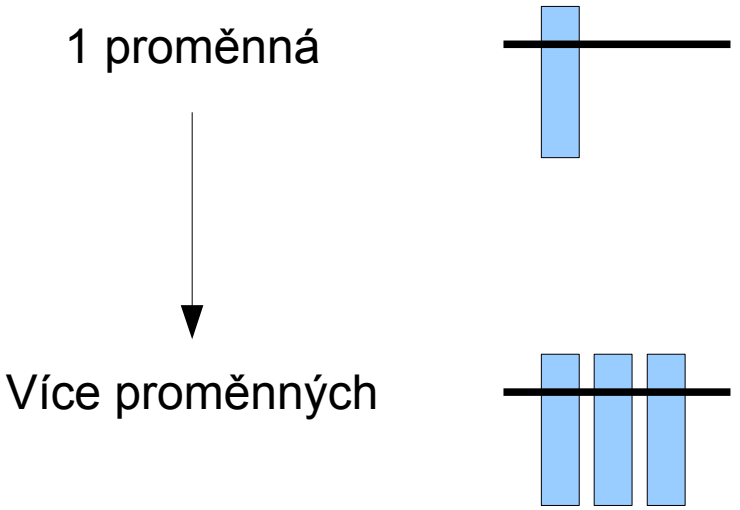
Relative warps Principal components of partial warp scores, sometimes weighted to emphasize components of low or high bending energy (that weighting is done by setting the parameter α to a value other than 0). Originally, the term referred to an eigenanalysis of the variance–covariance matrix relative to the bending-energy matrix, hence a new term was coined for these components (Bookstein, 1991). Currently, the term usually refers to a conventional principal components analysis of partial warp scores. See also ***Alpha*** (α), **Bending energy, Partial warp scores, Principal components analysis** (Chapter 7).

PCA - závěr

- PCA modeluje vztahy mezi landmarky pouze lineárně
  - Existují metody, které dokáží zachytit nelineární vztahy (dva landmarky se pohybují proti sobě, první s mocninou výchylky druhého, apod.)
- Další užitečné použití
  - Dobře separuje třídy – odlišnost se projevuje v prvních komponentách → klastrová analýza, CVA
  - Redukce dimenze – vytvoří stejný počet „nových“ proměnných, ale poslední nesou jen minimum informace, špatná interpretace významu proměnných
  - Dopočítávání chybějících dat

Statistika - obsah

- Analýza dat
- Základní pojmy z teorie pravděpodobnosti a statistiky
- Statistické testy
  - T-test, Hottelinguv test
  - Permutační testy
- Regresní analýza
- ANOVA, MANOVA
- Diskriminační analýza
- Shluková analýza



Doporučený software

- Past – PAlaeontological STatistics
  - tabulkový editor, dokumentace
  - Nabídka „Statistics“ - základní testy
  - Nabídka „Multivar“ - multivariační analýzy
  - Nabídka „Model“ - regresní analýza
- The R Project for Statistical Computing
  - Práce formou dialogu / psaní skriptů (scénářů, receptů)
  - Cokoliv...
- Matlab/Octave
- Excel
  - RExcel

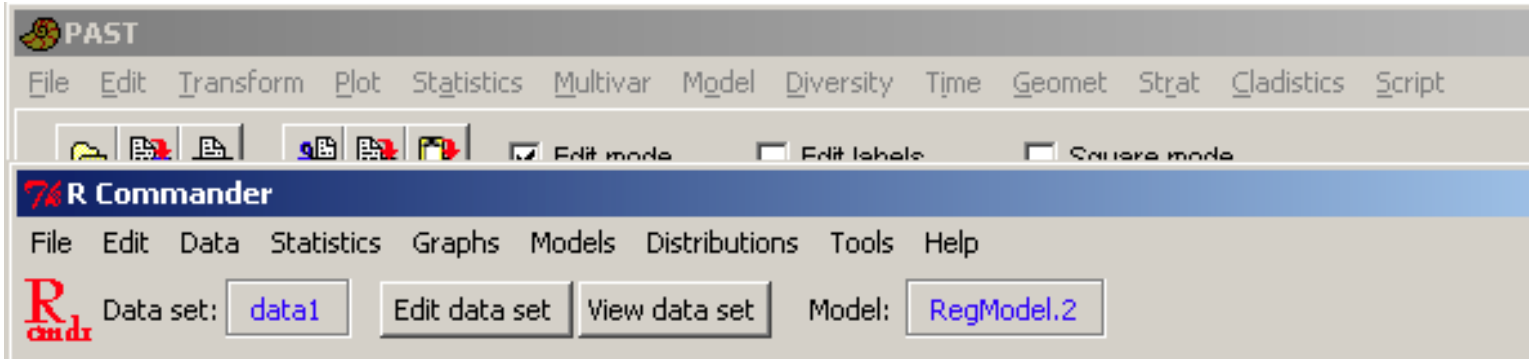
Past a R (R Commander)

- Podobné prostředí jako Past nabízí R rozšířený o modul R Commander (Rcmdr)
- Instalace

```
install.packages("Rcmdr", dependencies=TRUE)
```

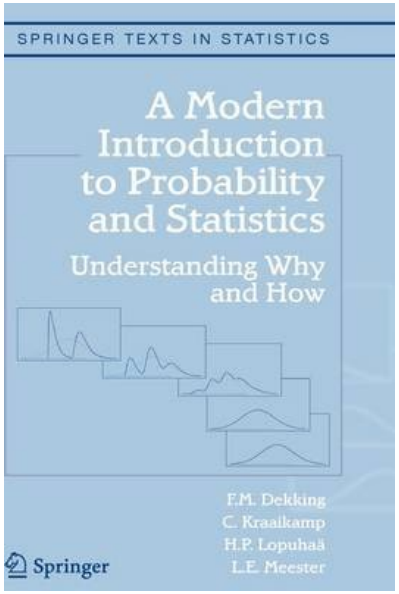
- Spuštění

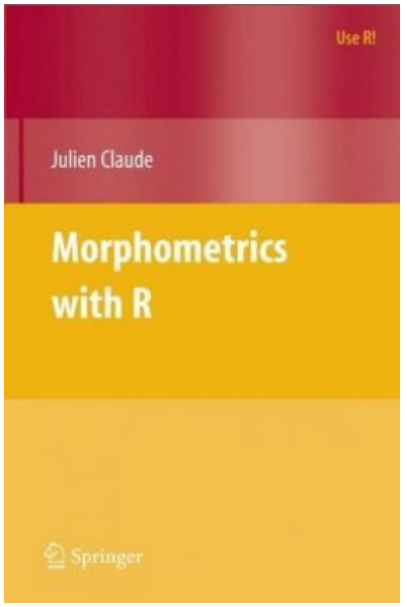
```
library(Rcmdr)
```

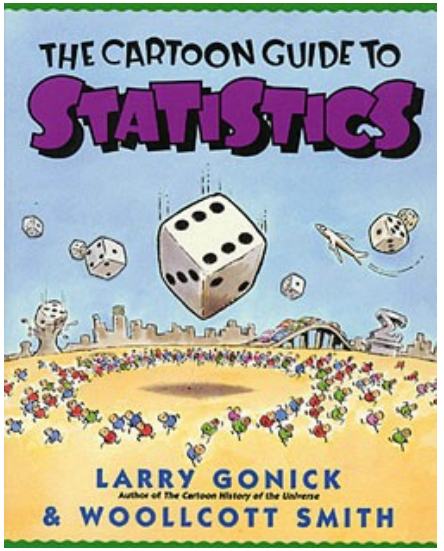


Doporučená literatura

- **http://cgg.mff.cuni.cz/~vajicek/gmm**
- GMM příručky z přednášky 1
- Dekking et al.: A Modern introduction to probability and Statistics
- Meloun, Militký: Kompendium statistického zpracování dat
- Claude: Morphometrics with R
- Gonick, Smith: Cartoon guide to statistics



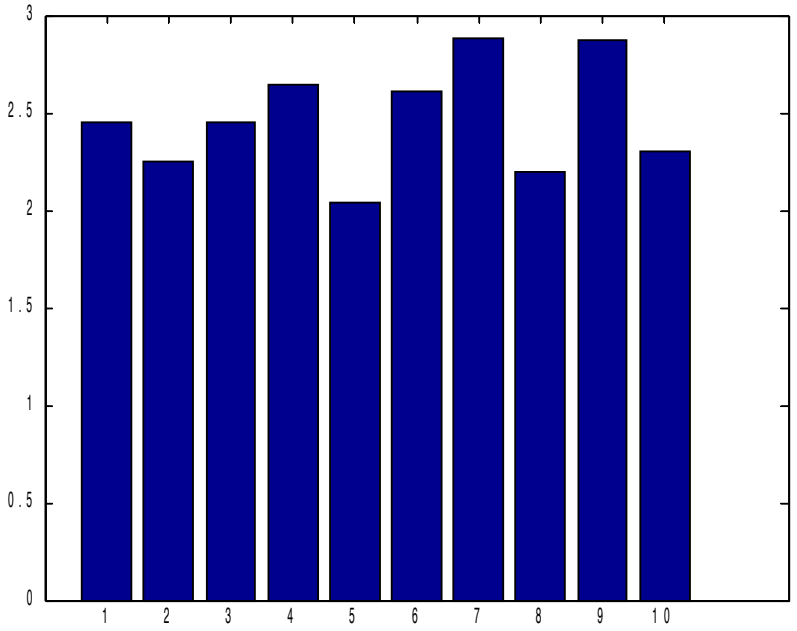


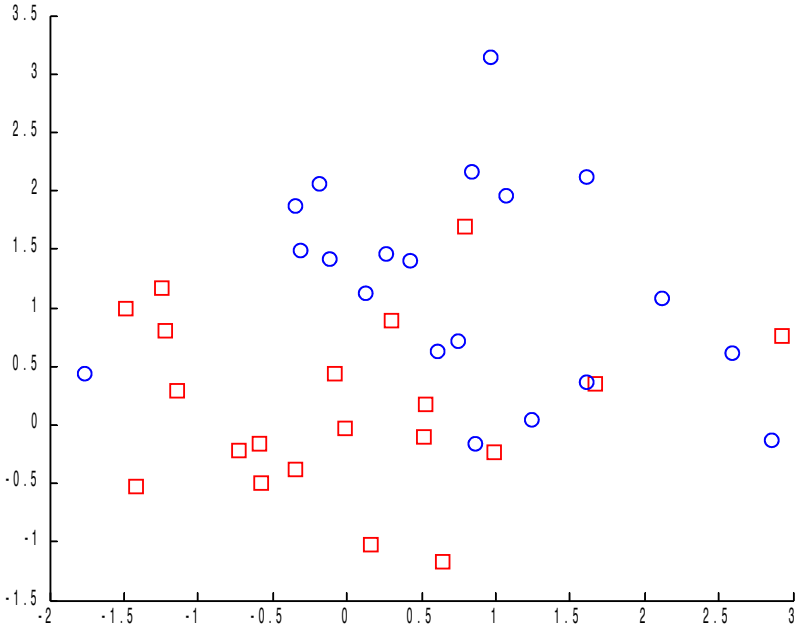


Analýza dat

1)Vizualizace hrubých naměřených dat

- Grafy
- Scatterplot





Analýza dat

2)Souhrny dat

- Průměr, směrodatná odchylka, median
- Kvantily
- Histogram

Naměřená data:

0.710 0.858 0.269 0.863 0.684 0.038 0.982 0.555 0.746 0.865 0.343

Setříděná data:

0.038 0.269 **0.343** 0.555 0.684 **0.710** 0.746 0.858 **0.863** 0.865 0.982

Q1 Median Q3

Jedna z možností výpočtu!

Pravděpodobnost

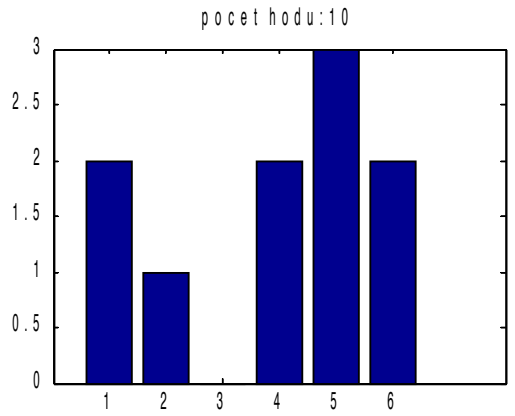
- Náhodný jev
  - Hod mincí, hod kostkou
  - Pravděpodobnost náhodného jevu
- Náhodná veličina
  - „Funkce na množině elementárních jevů“
  - Přiřazení čísel jevům
- Rozložení pravděpodobnosti
  - Popis náhodné veličiny
  - Různé modely rozložení/vzorečky
  - Odpovídá histogramu pro mnoho opakování

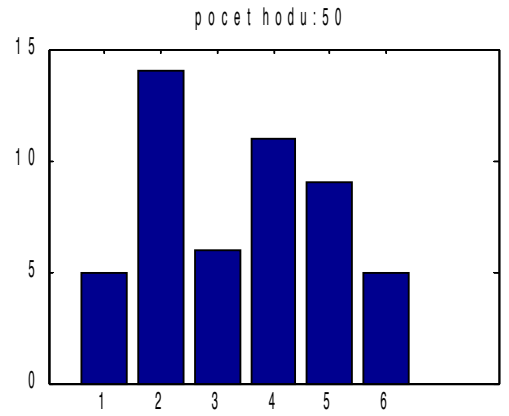
50 % panna $P_0 = 1/2$

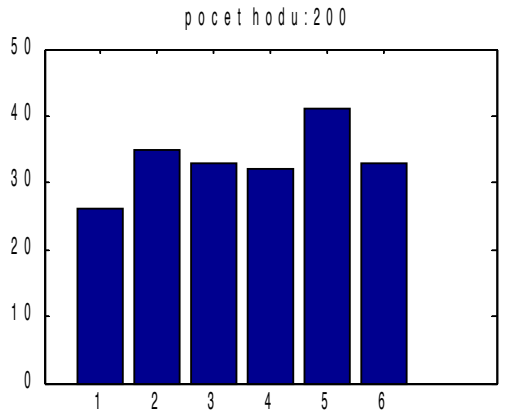
50 % orel $P_1 = 1/2$

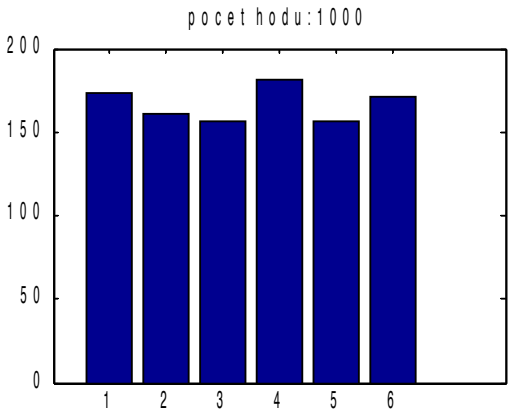
Experiment

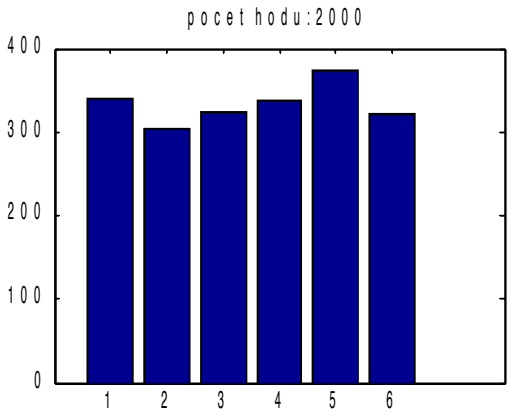
- Hod kostkou (k6)
  - Existují k8, k10 i k20, Proč?

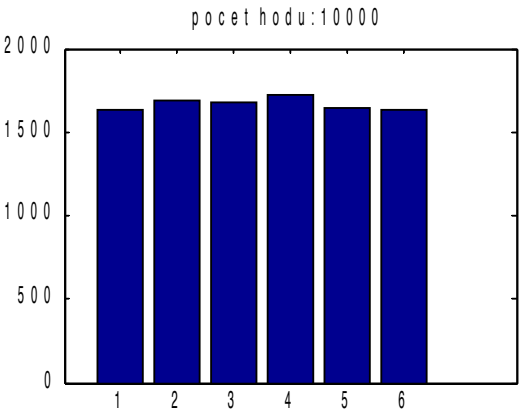






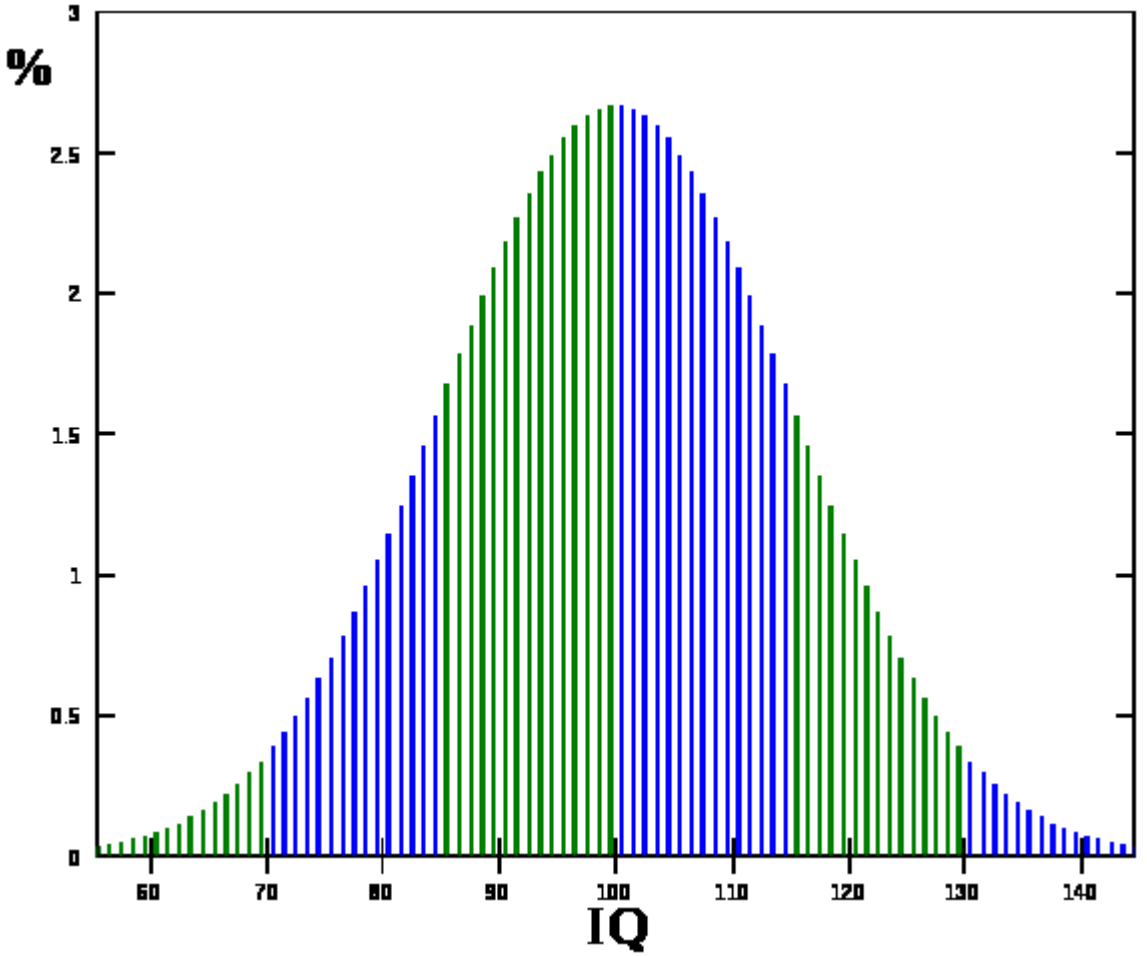






Příklad

- Inteligenční kvocient
  - Místo hodů kostkou budeme měřit kolemjdoucím IQ



Popis náhodné veličiny

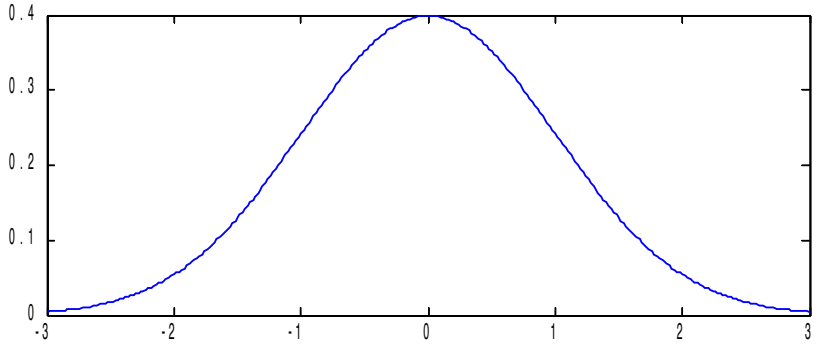
- Spojitá a diskrétní
- Střední hodnota ~ průměr
  - „Nejočekávanější hodnota“
- Rozptyl
  - Výběrový rozptyl
- Vyšší momenty
  - Šikmost, Špičatost

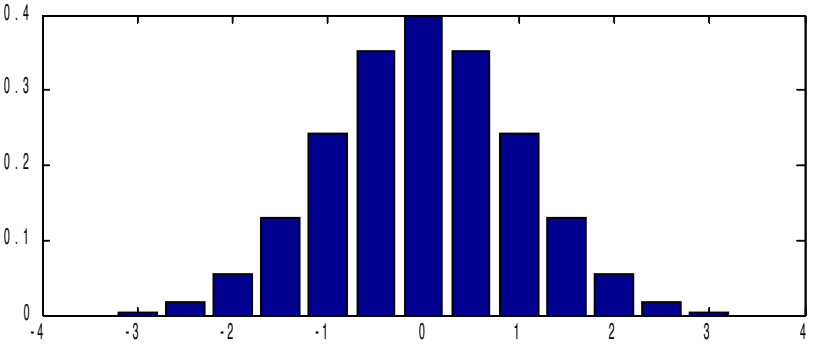
Pravděpodobnost všech jevů je stejná, tzv. **rovnoměrné rozložení**

$$\bar{x} = \frac{1}{N}\sum x_i = \sum p_i x_i \qquad E[X] = \int_{\infty}^{\infty} p(x)\,x\,dx$$

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{N} \qquad var(X) = E[(X - E(X))^2]$$

$$\sigma_n^2 = \frac{1}{N-1}\sum (x_i - \bar{x})^2$$





Normální rozložení

- *„Velké odchylky od očekávání jsou málo časté/pravděpodobné“*

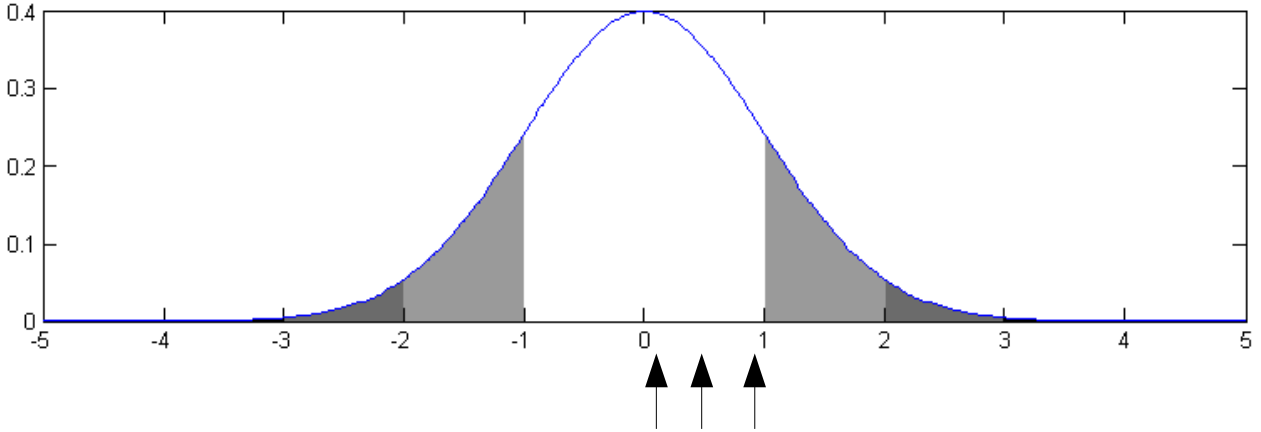
$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(X-X_0)^2}{2\sigma^2}}$$

- Součty libovolného rozložení mají normální rozložení
  - Centrální limitní věta - komplexní děje se skládají z mnoha náhodných událostí → normální rozložení je všude
- Z-transformace
  - Posunutí a zúžení rozložení, aby byla výsledná střední hodnota 0 a střední kvadratická chyba 1

$$Z = \frac{X-\mu}{\sigma(X)}$$

Normální rozložení

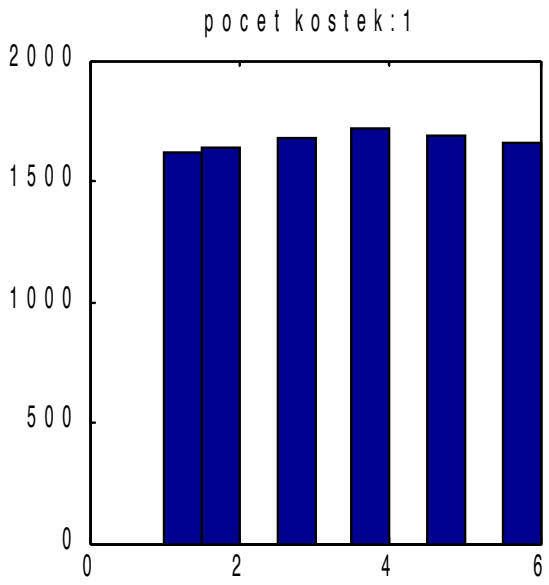
- Z-skóre
  - Tabulka rozložení, symetrie rozložení
  - Celková plocha pod grafem = 1

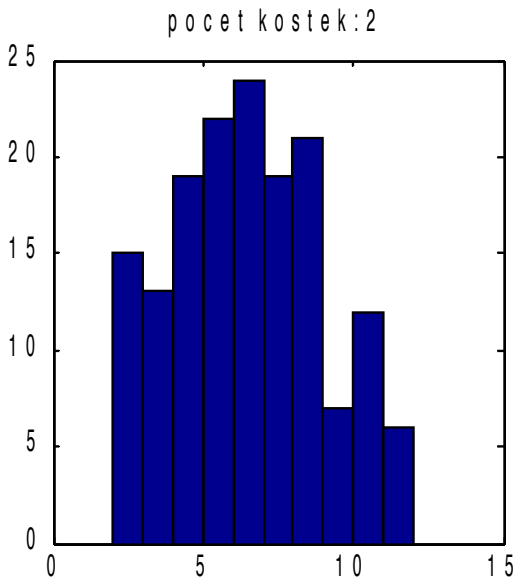


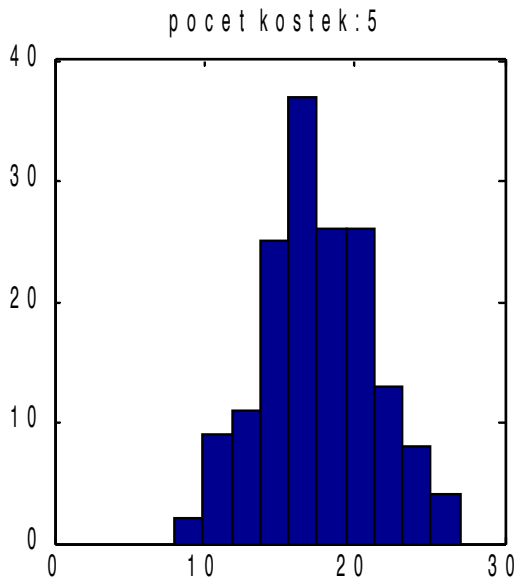
Z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879

Příklad

- Hod nk6



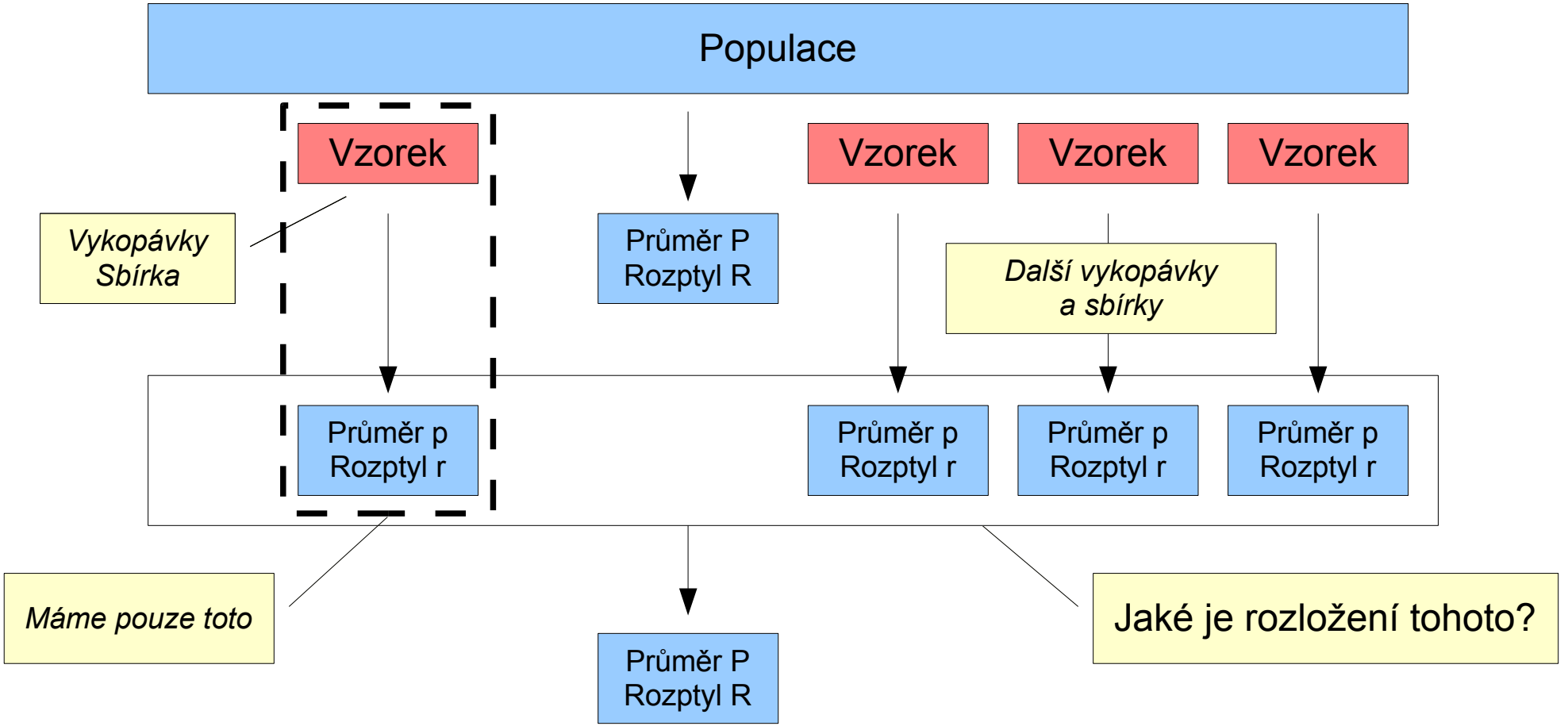




Populace a vzorek

- Chceme studovat celou populaci, ale pracujeme jen se vzorkem → **deskriptivní statistika**
- Pozorování na vzorku zobecnit na populaci → **inferenční statistika**
  - Průměr na vzorku není identický s průměrem populace
  - Průměr jednoho vzorku je náhodný jev
  - Průměr populace je střední hodnota náhodné veličiny
  - *Závěry platí s jistou pravděpodobností (závisí na velikosti vzorku, způsobu výběru vzorku)*
- V GMM pracujeme se vzorky a ne celou populací → důvod proč zkoumat jaký dopad mají naše zjištění
- I měření jednotlivců se dá použít inference

Populace a vzorek



t-rozložení

- Rozložení pravděpodobnosti rozptylu při výběru vzorku z populace
- Vzorek je malý a neznáme rozptyl populace
  - Použijeme rozptyl vzorku (výběrový rozptyl)
- Parametrizované velikostí vzorku
  - Vhodné pro GMM kde se většinou pracuje s malými vzorky
- Komplikovaný výpočet funkce
  - Použití tabulky
- t-hodnota

$$\frac{X-\mu}{\sigma_n/\sqrt{n}} \rightarrow \frac{X-\mu}{s/\sqrt{n}}$$

	0,100	0,050	0,025	0,010	0,005	0,001	0,0005
1	3,078	6,314	12,710	31,820	63,660	318,300	637,000
2	1,886	2,920	4,303	6,965	9,925	22,330	31,600
3	1,638	2,353	3,182	4,541	5,841	10,210	12,920
4	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	1,476	2,015	2,571	3,365	4,032	5,893	6,869
6	1,440	1,943	2,447	3,143	3,707	5,208	5,959
.	...	...	...	...	...	...	...

Test hypotézy

- Ověřit pravdivost nějakého tvrzení o datech
  - Rovnost/nerovnost dvou vzorků
  - Rovnost/nerovnost střední hodnoty konkrétní hodnotě
- Jediný důkaz pro podporu/vyvrácení je v datech
- Postup
  - Nulová hypotéza(H0) : „X má střední hodnotu 0“
  - Alternativní hypotéza(Ha): „X nemá střední hodnotu 0“, ...
  - Určení *skóre* jevu popírajícího nulovou hypotézu
  - Výpočet p-hodnoty, pravděpodobnosti že pozorovaný jev je dílem náhody
  - Porovnání s kritickou hodnotou (nejčastěji 0.05, 0.1)

obtížné

Příklad – jednovýběrový 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]
```

- H0: Data mají střední hodnotu 0
- Vzorků je málo → použiji t-rozložení a t-hodnotu

$$t = \frac{\bar{X} - \mu_0}{SE(\bar{X})} \qquad t = \frac{1.345 - 0}{0.931} = 1.445$$

- Porovnání t-hodnoty s kritickou hodnotou t-rozložení
  - Vyvrácení H0 (přijmutí H0)

Počet stupňů volnosti je n-1, jeden stupeň je střední hodnota vzorku

	0,100	0,050	0,025	0,010	0,005	0,001	0,0005
.	...	...	...	...	...	...	...
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
.	...	...	...	...	...	...	...

Porovnání dvou vzorků

- Dva vzorky z dvou populací → signifikantní rozdíl?
- Vzorek každé populace je náhodná veličina

$$X_1 \qquad\qquad X_2$$

- Rozdíl středních hodnot je také náhodná veličina

$$\mu_1 - \mu_2$$

- Jaké je rozložení, střední hodnota, rozptyl?
  - Pro **velké vzorky** je rozložení normální → kritická hodnota
  - Kritická hodnota → výpočet odhadu střední hodnoty

$$\mu_1 - \mu_2 = \bar{\mu}_1 - \bar{\mu}_2 \pm z\, SE(\bar{X}_1 - \bar{X}_2)$$

Porovnání dvou vzorků

- Jak spočítat směrodatnou odchylku?
  - Za předpokladu **normality** a **nezávislosti**

$$SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma^2(\bar{X}_1)}{n_1} + \frac{\sigma^2(\bar{X}_2)}{n_2}}$$

- Porovnání jako test hypotézy
  - Vzorky stejnou střední hodnotu → rozdíl nulovou

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{SE(\bar{X}_1 - \bar{X}_2)}$$

= 0

Výpočet z-skóre (statistiky) daného vzorku

Dvouvýběrový t-test

- Pro **malé vzorky** se používá statistika t-skóre
  - Předpoklad **normality** a **nezávislosti**
  - Neznámý ale **stejný rozptyl**
- Pro odhad směrodatné odchylky kombinujeme rozptyly vzorků

$$SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{(n-1)\sigma^2(\bar{X}_1) + (m-1)\sigma^2(\bar{X}_2)}{n+m-2}\left(\frac{1}{n}+\frac{1}{m}\right)}$$

- P-hodnotu určím přibližně pomocí tabulky t-rozdělení s n+m-2 stupni volnosti

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{SE(\bar{X}_1 - \bar{X}_2)}$$

Co se nestihlo

MFF UK Prague

- Regresní analýza
- ANOVA
  - One-way ANOVA
  - Two-way ANOVA
- MANOVA
- Hotellingův test
- Diskriminační analýza
- Shluková analýza