Stochastické modelování

Jiří Jarkovský

Organizace kurzu

- Každých dní (úterý 8-12 hodin) – počítačová učebna IBA

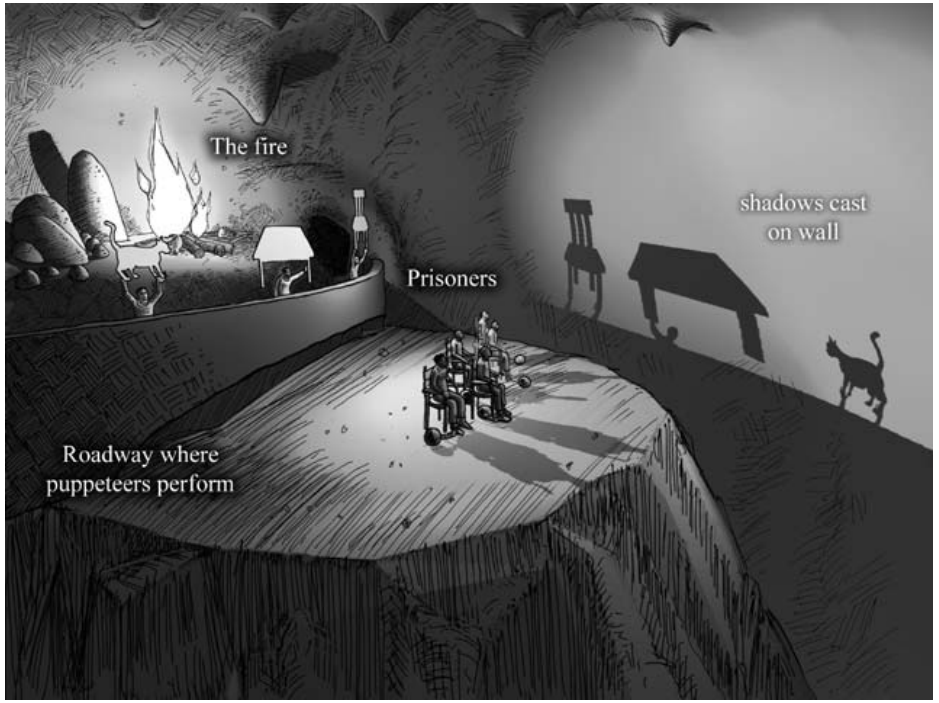
- Obecný úvod do problematiky modelování
  - Principy
  - Statistické hodnocení dat
  - Vícerozměrná analýza
- „Klasické" statistické modelování
  - ANOVA
  - Regrese
  - GLM, GAM
- Pokročilé metody
  - Prostorové modelování
  - Rozhodovací stromy a lesy

Statistika: o čem to vlastně je?

Výzkum, reálný svět a statistika

- ☑ **Výzkum je způsobem poznávání světa**
- ☑ **Jak přesné a pravdivé je naše pochopení světa?**

Statistika je jedním z nástrojů přinášejících spolehlivost do našich výsledků



Variabilita našeho světa

- **Varabilita je základem našeho světa a statistika je vědou zabývající se variabilitou**
- **Správná analýza a vysvětlení variability nám dává informaci o světě**
- **V případě deterministického světa je statistika zbytečná**

Statistika

☑ Co tento termín znamená?

WWW.WIKIPEDIA.ORG:

Statistics is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data. It is applicable to a wide variety of academic disciplines, from the physical and social sciences to the humanities. Statistics are also used for making informed decisions and misused intentionally or accidently.

Statistika používá matematické modely reality pro zobecnění našich informací získaných z experimentů a vzorkování.
Statistika je korektní pouze při naplnění předpokladů jejích matematických modelů !!!

Smysl zpracování dat

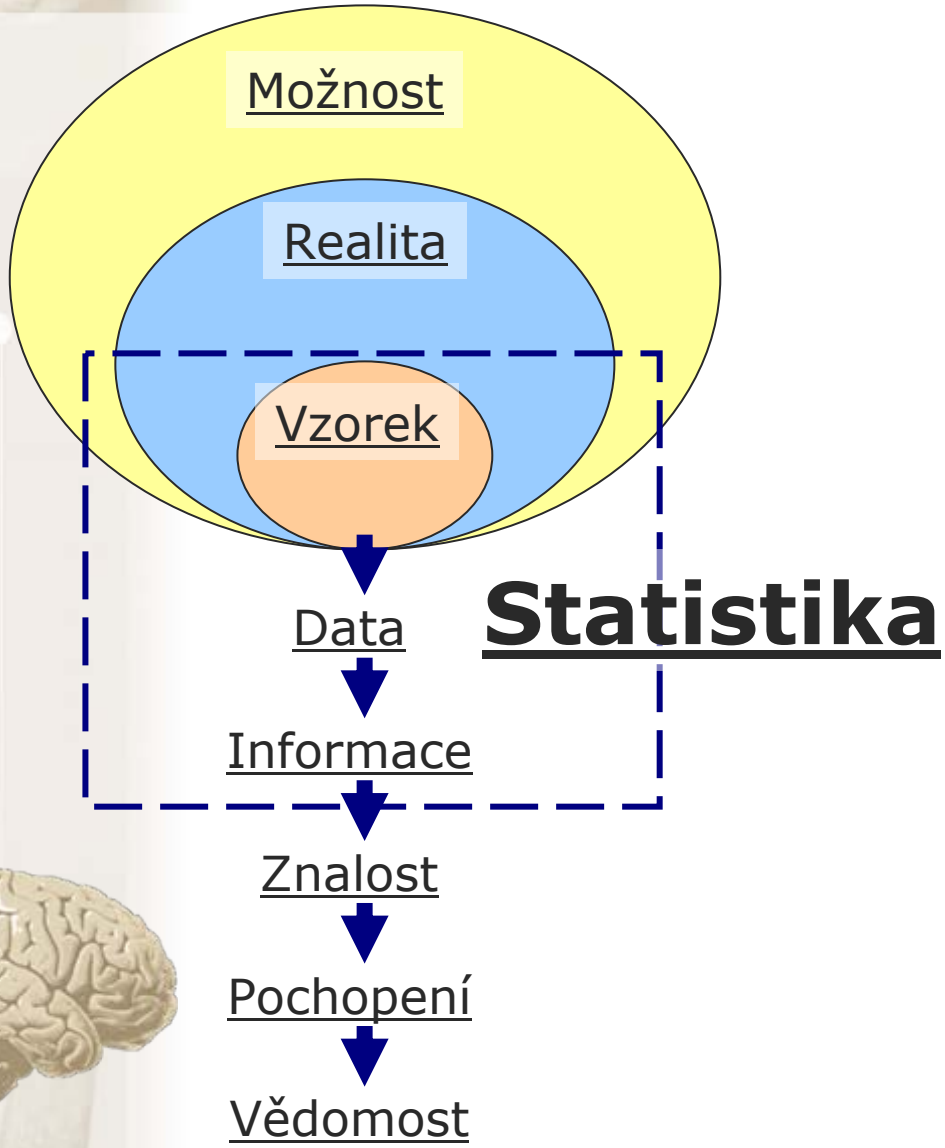
CO JE CÍLEM

- ☑ Srozumitelný popis dat a jejich vizualizace
- ☑ Porozumění vztahům mezi proměnnými
- ☑ Získat znalost v dané oblasti
  - ➔ Schopnost interpretace
  - ➔ Využití této znalosti dále
- ☑ Korektní postup při zpracování

CO NENÍ CÍLEM

- ☑ Pouhý přehled čísel
- ☑ Vzorce, rovnice a matematickou teorii
- ☑ Neinterpretovatelné výsledky
  - ➔ Rozpor mezi zkušeností, selským rozumem a čísly
  - ➔ Obvykle znamenají chybu

Co říká statistika o našem světě?



Statistika neříká nic o věcech nespojených s naším vzorkem.

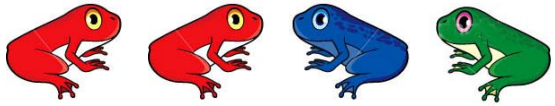
Statistika může být využita při **získání informací z vzorkovaných dat** a jako podpora naší znalosti a pochopení problému.

Statistika není náhradou naší inteligence !!!

Vzorkování ve statistickém slova smyslu

- **Statistika hovoří o realitě skrz vzorek !!!**
  - **Statistické předpoklady korektního vzorkování.**

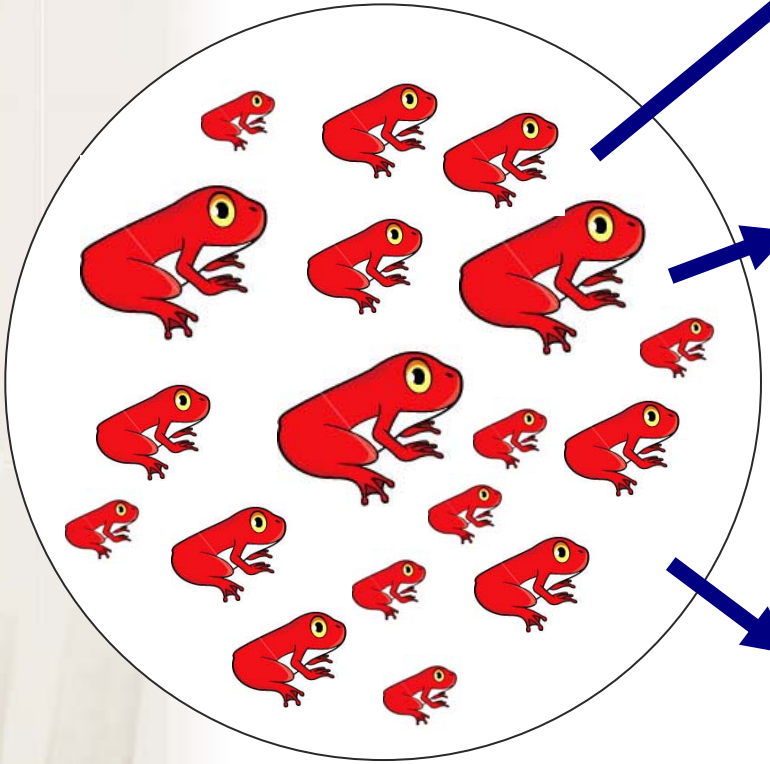
Reprezentativnost: struktura vzorku by měla kopírovat realitu jak je to jen možné



Nezávislost: v opakovaném vzorkování stejného objektu není žádná nová informace

Velikost vzorku a přesnost statistiky

Existuje **skutečné rozložení** a **skutečný průměr** proměnné.



Z jednoho měření nevíme nic

Vzorek:  → **?????**

Vzorek určité velikosti poskytuje **odhad skutečné hodnoty** s určitou **spolehlivostí**.

vzorek:  → **Odhad průměru, SD atd.**

Vzorkování všech hodnot poskytne **skutečnou průměrnou hodnotu** proměnné, nicméně je v realitě většinou **nemožné**.

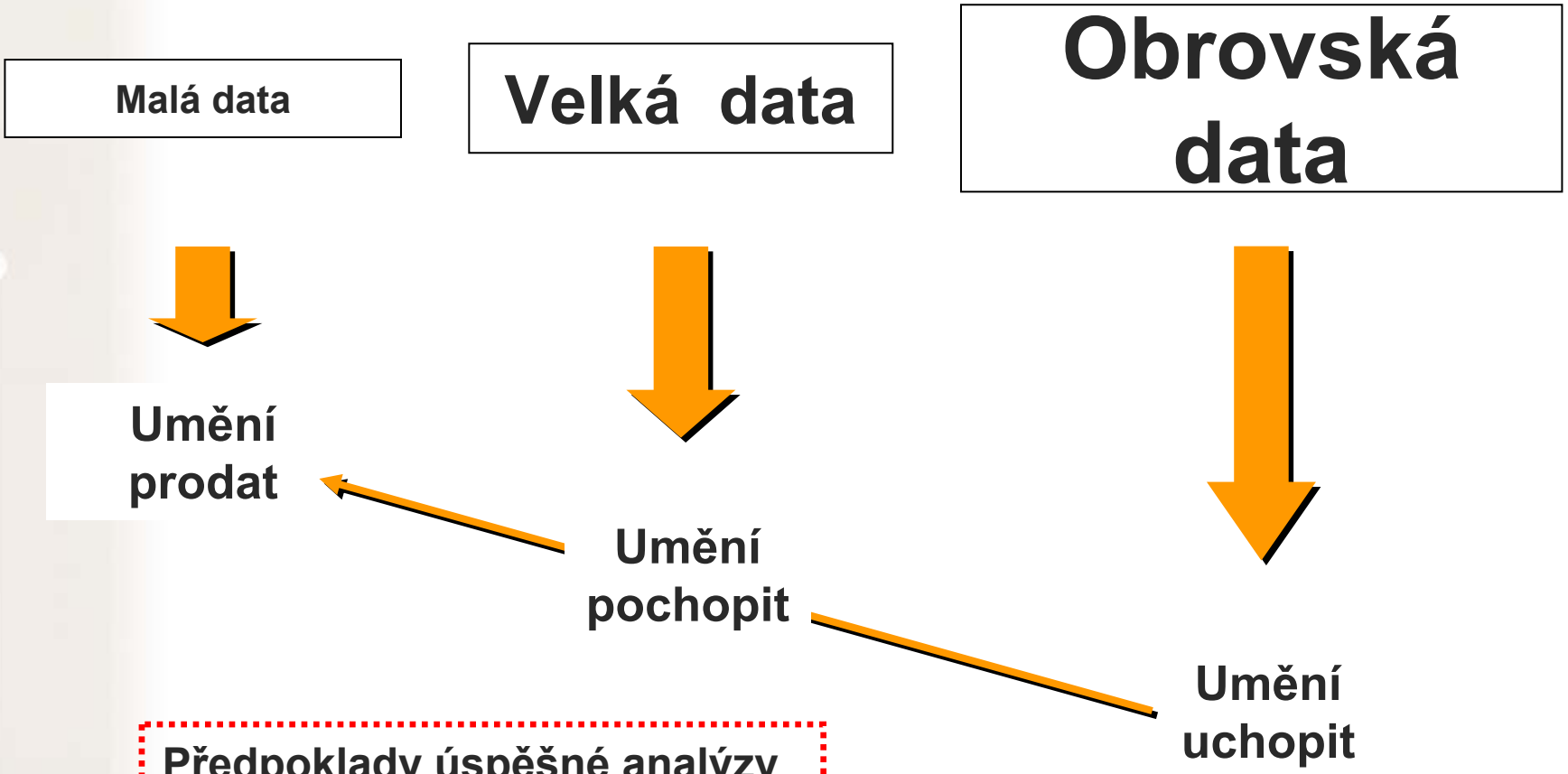
Data a jejich popis

Jak vznikají data?

Záznamem skutečnosti …

… více či méně dokonalým

Charakter dat



Předpoklady úspěšné analýzy

- Znalost problému
- Znalost analýzy dat

Typy dat

Data poměrová

Kolikrát ?

Hladina cholesterolu v krvi, Doba do progrese

Data intervalová

O kolik ?

Teplota ve °C

Spojitá data

Data ordinální

Větší, menší ?

Kategoriální otázky

Performance status, Počet mobilních telefonů

Data nominální

Rovná se ?

Otázky „Ano/Ne“

Pohlaví, Krevní skupina

Diskrétní data

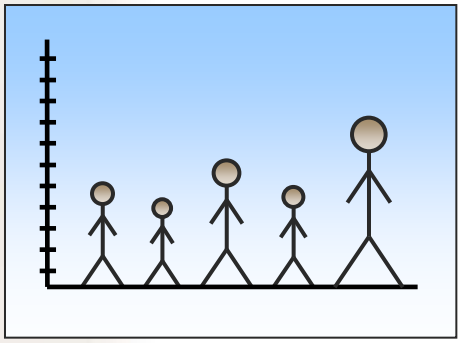
Samotná znalost typu dat ale na dosažení informace nestačí .....

Data = záznam informace

- **Různé typy dat sumarizujeme a analyzujeme různým způsobem → jiná úskalí při analýze**
- **Různé analytické metody jsou vhodné pro různé typy dat**

- **Data jsou variabilní**
  - **Vysvětlení variability (např. v odpovědi pacienta na léčbu) je naším hlavním cílem**
  - **Bez variability by nebylo potřeba analýzy dat**

Data jsou variabilní



VARIABILITA

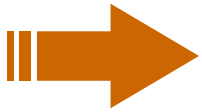
?

CHYBA **INFORMACE**

Cílem analýzy dat je popis a vysvětlení maximálního množství variability v datech, zbytek lze přisuzovat chybám měření.

Variabilita dat?

Variabilita opakovaných měření

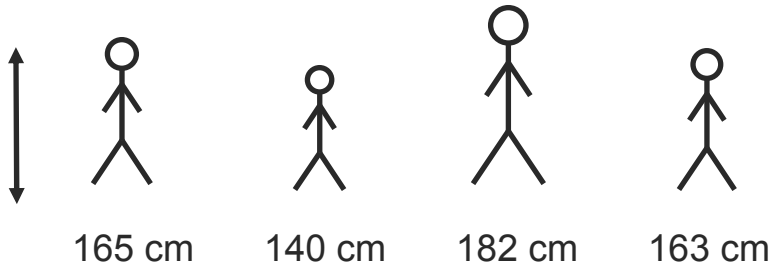


Data

2,1
2,8
3,2
1,2
5,2
2,9

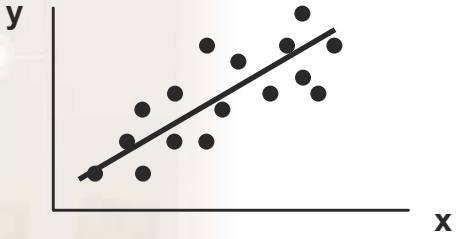
chyba

Variabilita znaku v populaci



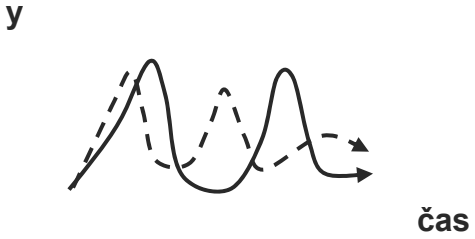
rozptyl znaku, přirozená variabilita

Variabilita modelovaných dat



chyba = nepřesnost modelu

Variabilita časových řad



fluktuace, časová proměnlivost

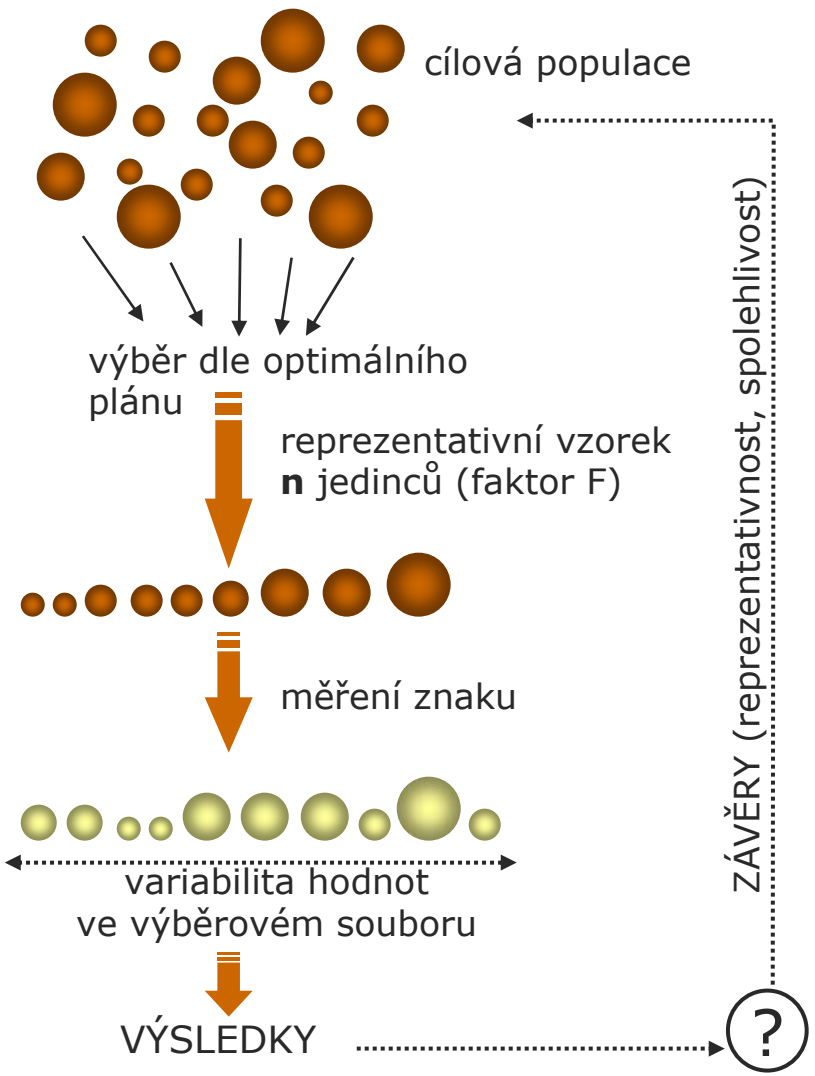
Variabilita ve skladbě biologických společenstev

DRUH 1	15
DRUH 2	30
DRUH 3	40
DRUH 4	14

biodiverzita

Variabilita – její popis

Účel analýzy: Popisný

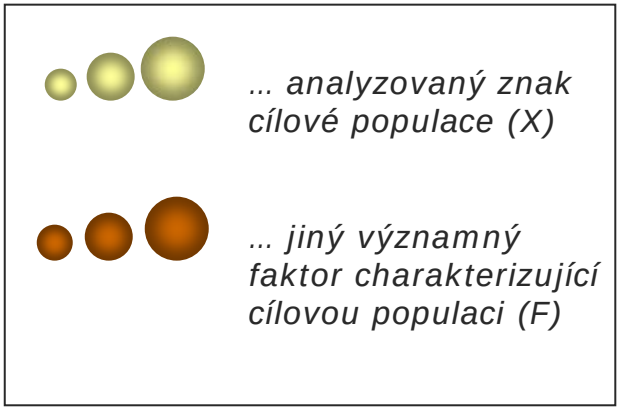


?

Reprezentativnost

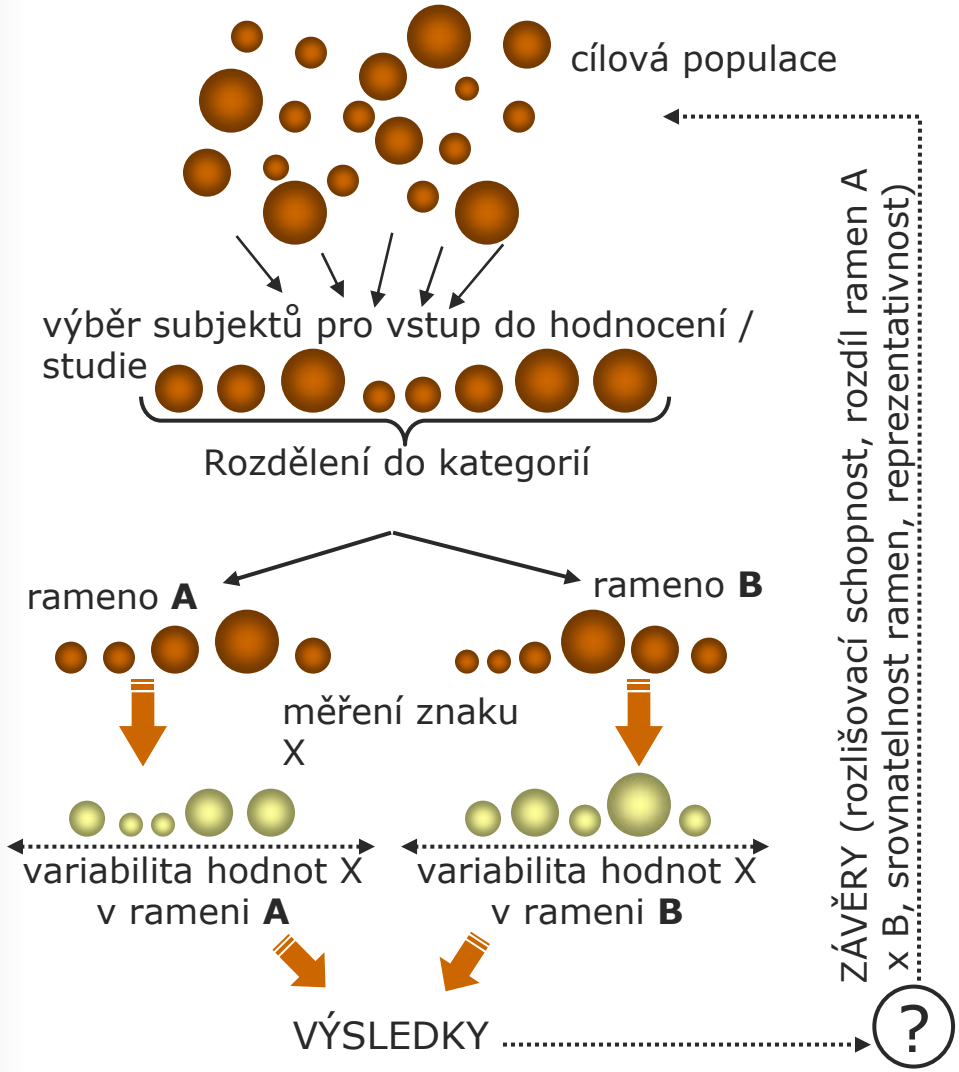
Spolehlivost

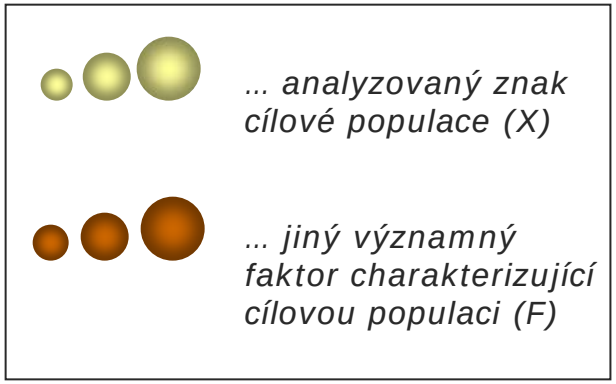
Přesnost



Variabilita – srovnání dvou skupin

Účel analýzy: Srovnávací





?

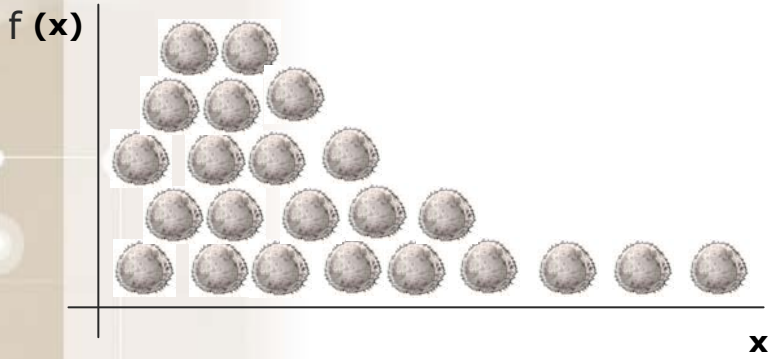
Srovnatelnost

Spolehlivost

Přesnost

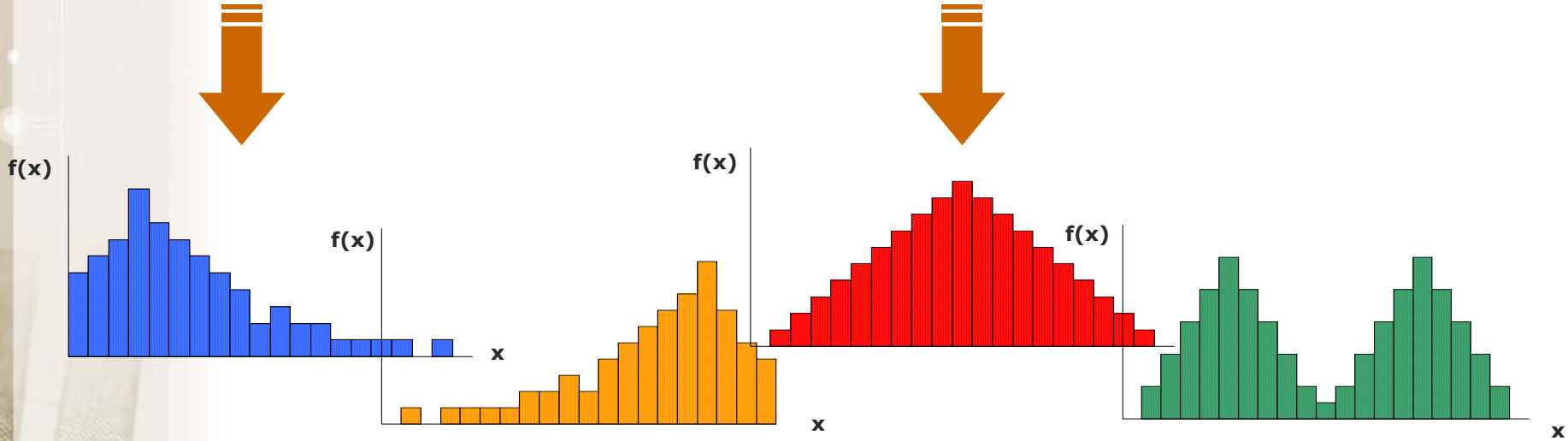
Rozdělení hodnot jako základ statistického hodnocení

Data podléhají určitému rozdělení hodnot

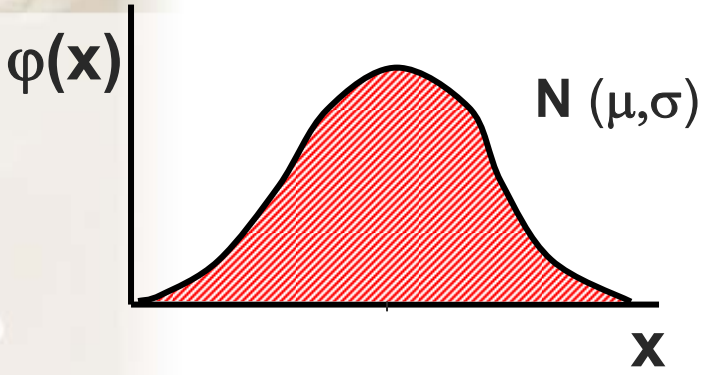


Počet bílých krvinek

Výška



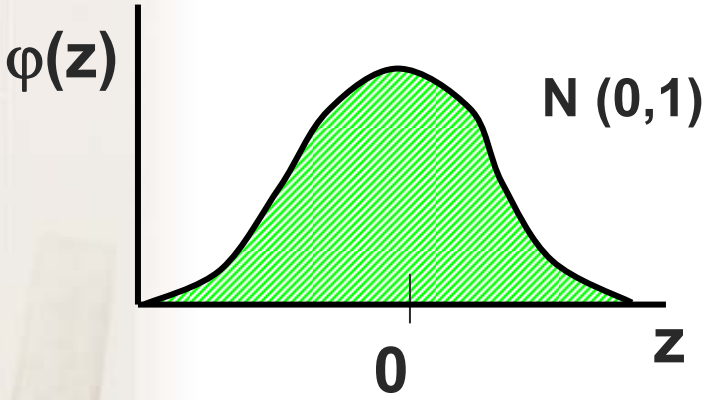
Rozložení hodnot jako model: Příklad - Normální rozložení



$$\varphi(x) = \frac{1}{\sigma.\sqrt{2\pi}}.e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$z = \frac{x - \mu}{\sigma}$$

Standardizovaná forma



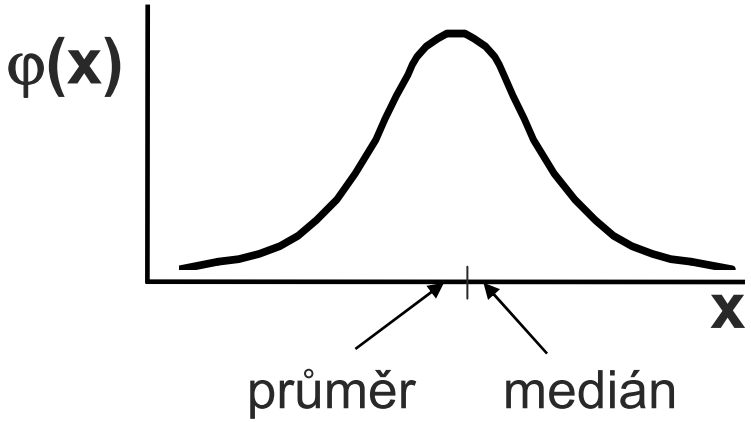
$$\varphi(z) = \frac{1}{\sqrt{2.\pi}}.e^{-\frac{z^2}{2}}$$

Tabelovaná podoba

Parametry charakterizující normální rozložení a jejich význam

$E(x) \sim \bar{x} \sim \mu$

$D(x) \sim s^2 \sim \sigma^2$



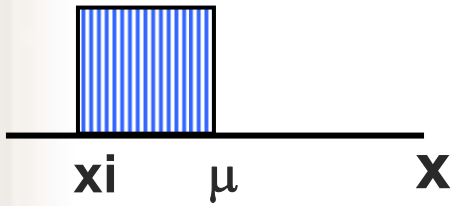
a) $\mu \sim \bar{x}$

průměr - ukazatel středu

b) $\sigma^2 \sim s^2$

rozptyl

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



c) $\sigma \sim s$

směrodatná odchylka

$$s = \sqrt{s^2}$$

Pravidlo ± 3s

d)

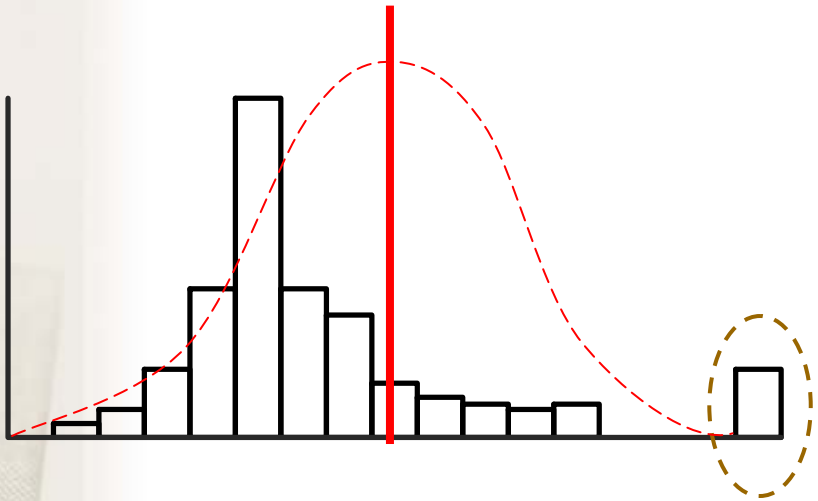
koeficient variance

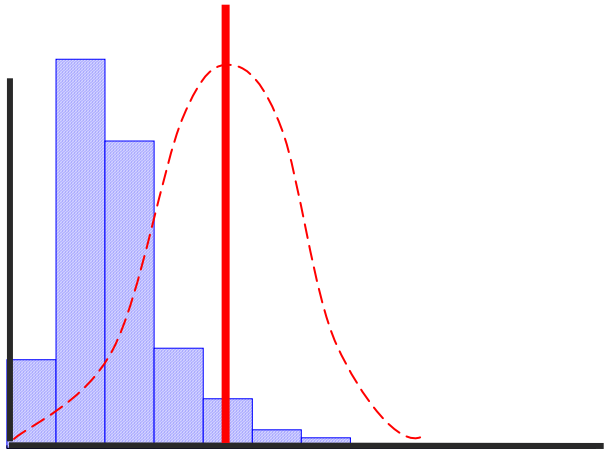
$$c = s / \bar{x}$$

Průměr jako odhad střední hodnoty

- ☑ Statistické výpočty jsou často postaveny na teoretických předpokladech
- ☑ Tyto předpoklady platí jak pro klasické, tak pro vícerozměrné analýzy
- ☑ Tyto předpoklady nejsou často dodrženy, zejména při malém vzorku
- ☑ Všechny výpočty založené na předpokladu normálního rozložení mohou poskytovat zavádějící výsledky
- ☑ Průměr jako ukazatel středu je silně ovlivněn tvarem dat a odlehlými hodnotami

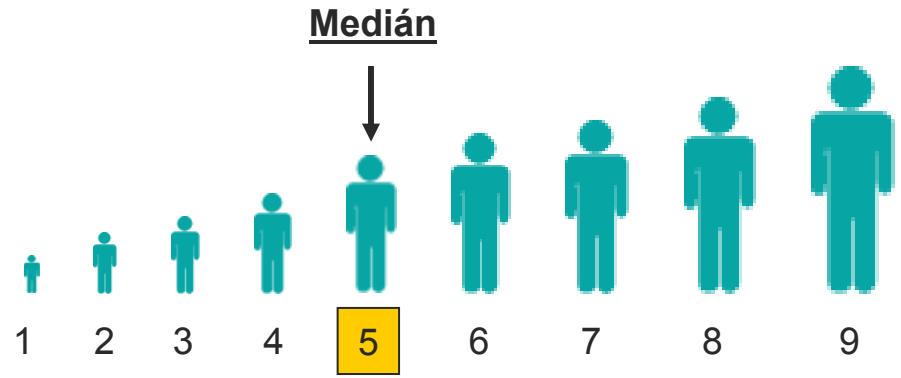
V případě odlehlých hodnot nebo asymetrie si analýza myslí, že existuje „normální“ tvar dat a průměr



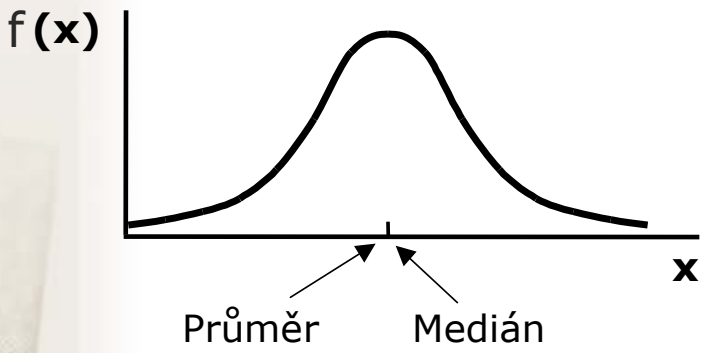


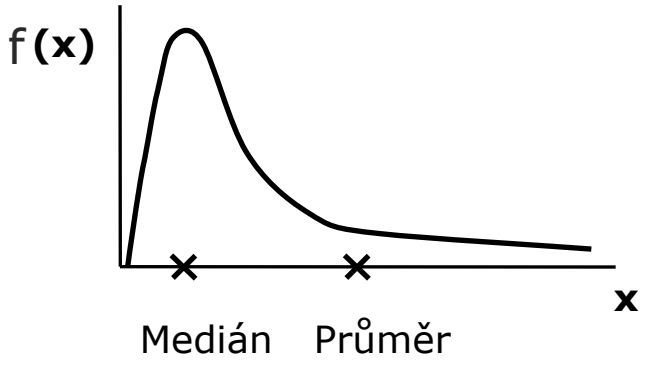
Popis středu dat – průměr a medián

- **Průměr** – vhodný ukazatel středu u normálního/symetrického rozložení, kde x_i jsou jednotlivé hodnoty a N jejich počet
- **Medián** – jde vlastně o 50% kvantil, tj. polovina hodnot leží nad a polovina pod mediánem



- V případě symetrického rozložení jsou jejich hodnoty v podstatě shodné





Průměr vs. medián

Příklad známkování ve škole:

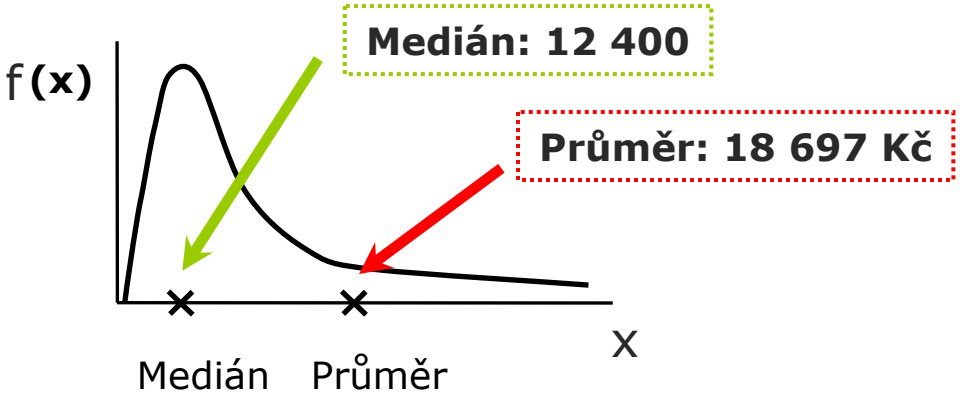
Student A: 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 5 → (N = 14)

Průměr: 1.35 **Medián: 1.00**

Student B: 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 2 → (N = 14)

Průměr: 1.13 **Medián: 1.00**

Příklad platu v ČR v roce 2003:



Odhady parametrů

Bodové

Číslo (chyba)
(Odhad parametru)

Intervalové

Interval pravděpodobných hodnot

Spolehlivost
(Pravděpodobnostní interpretace)

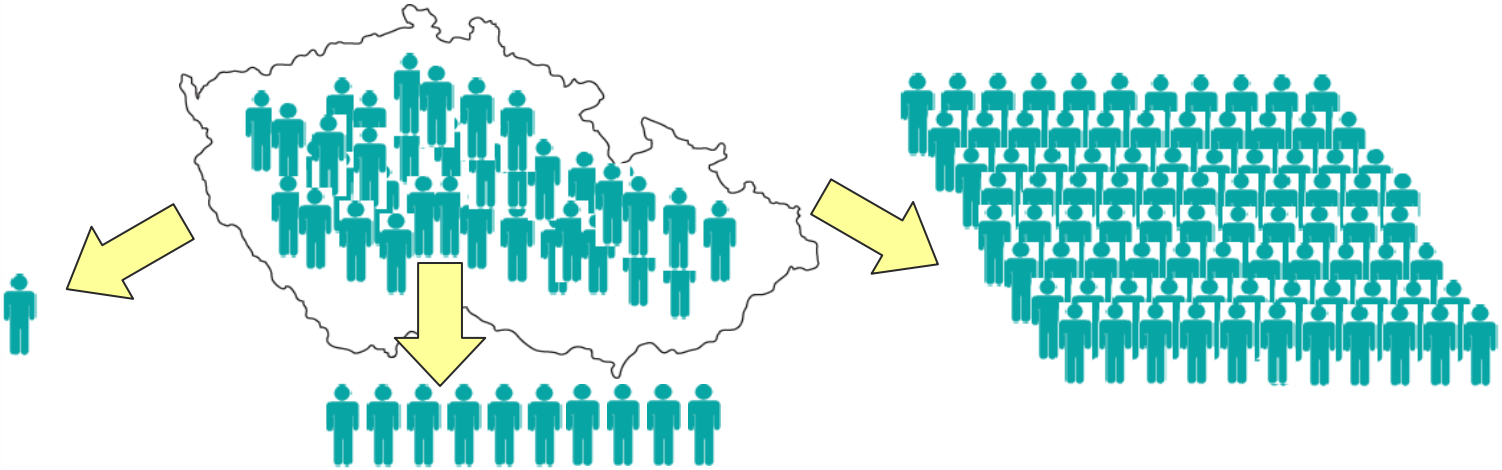
Obecný tvar: $P\ (L_1 < \text{Odhad} < L_2) \geq 1 - \alpha/2$

Odhadovaný parametr $\pm$ Kvantil modelového rozložení . SE (odhadu)

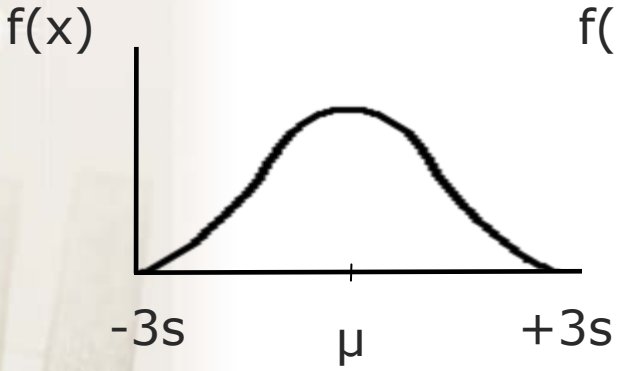
K_V pro $(1 - \alpha/2)$

Interval spolehlivosti - informace o přesnosti odhadu

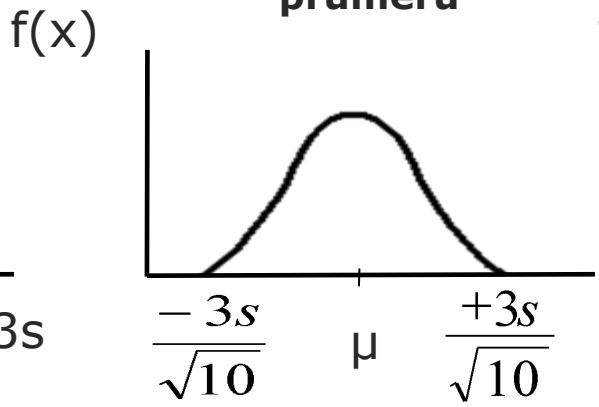
- Se zvětšující se velikostí vzorku (při zachování) reprezentativnosti se zvětšuje přesnost našeho odhadu o celém trhu
- Interval spolehlivosti je hodnocen pro (1 - α) procentní spolehlivost



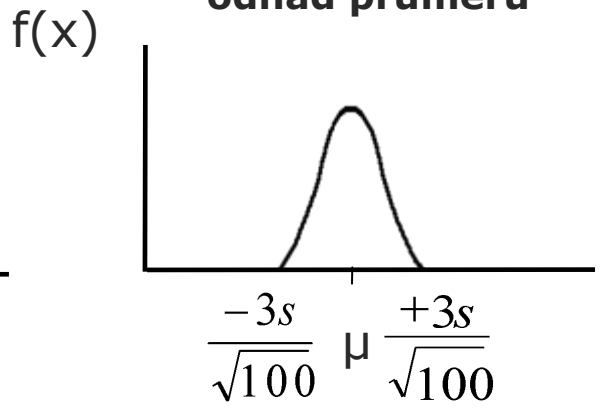
Původní proměnná x



Výběr n=10 pro odhad průměru



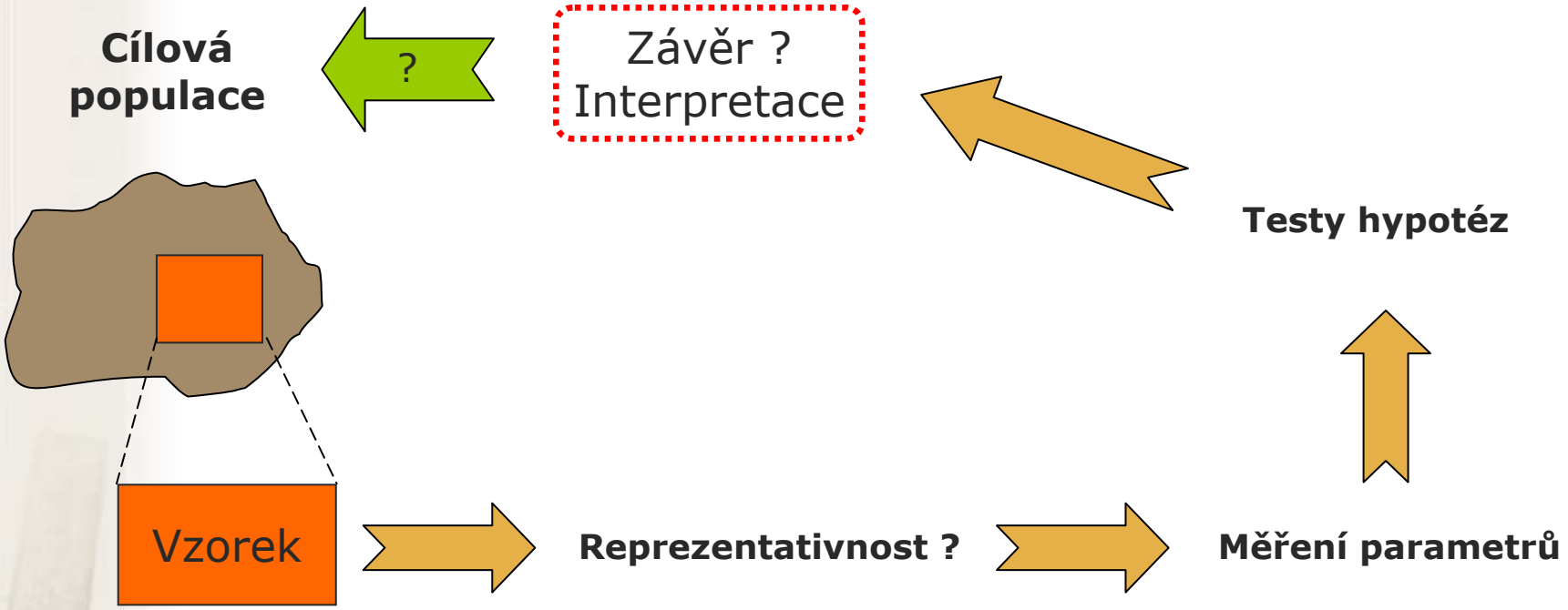
Výběr n=100 pro odhad průměru



Statistické testování

Princip testování hypotéz

- Formulace hypotézy
- Výběr cílové populace a z ní reprezentativního vzorku
- Měření sledovaných parametrů
- Použití odpovídajícího testu ⟹ závěr testu
- Interpretace výsledků

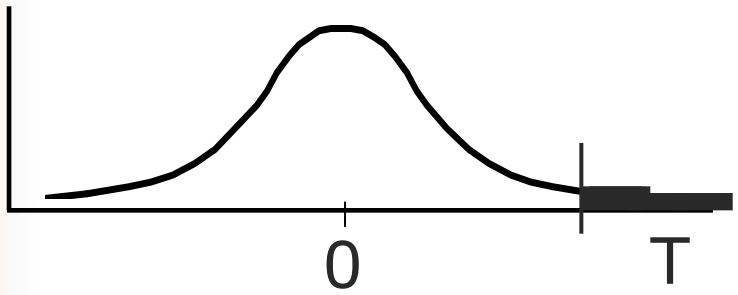


Statistické testování – základní pojmy

- **Nulová hypotéza H_0** — H_0: sledovaný efekt je nulový
- **Alternativní hypotéza H_A** — H_A: sledovaný efekt je různý mezi skupinami
- **Testová statistika**

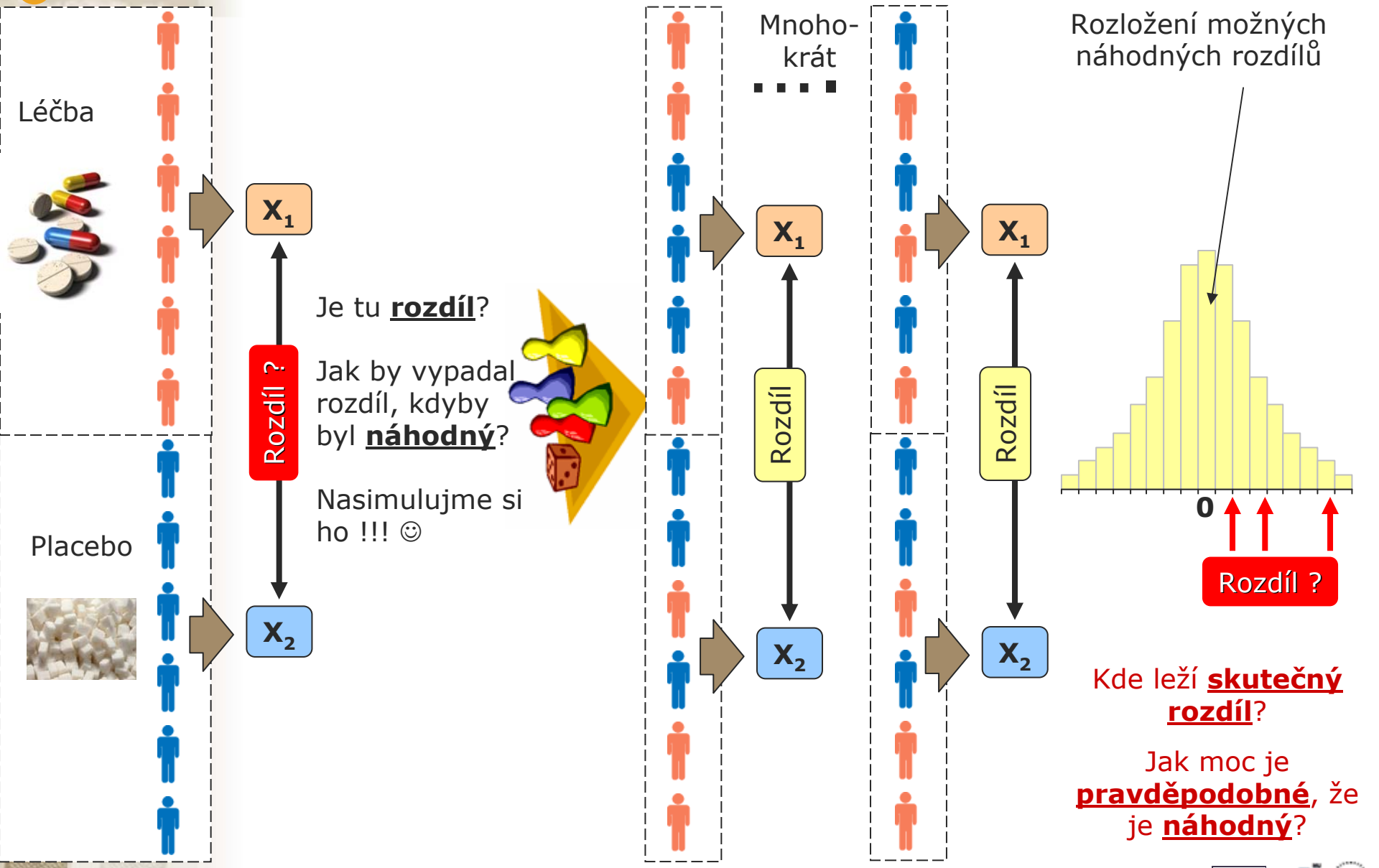
$$\text{Testová statistika} = \frac{\text{Pozorovaná hodnota} - \text{Očekávaná hodnota}}{\text{Variabilita dat}} * \sqrt{\text{Velikost vzorku}}$$

- **Kritický obor testové statistiky**



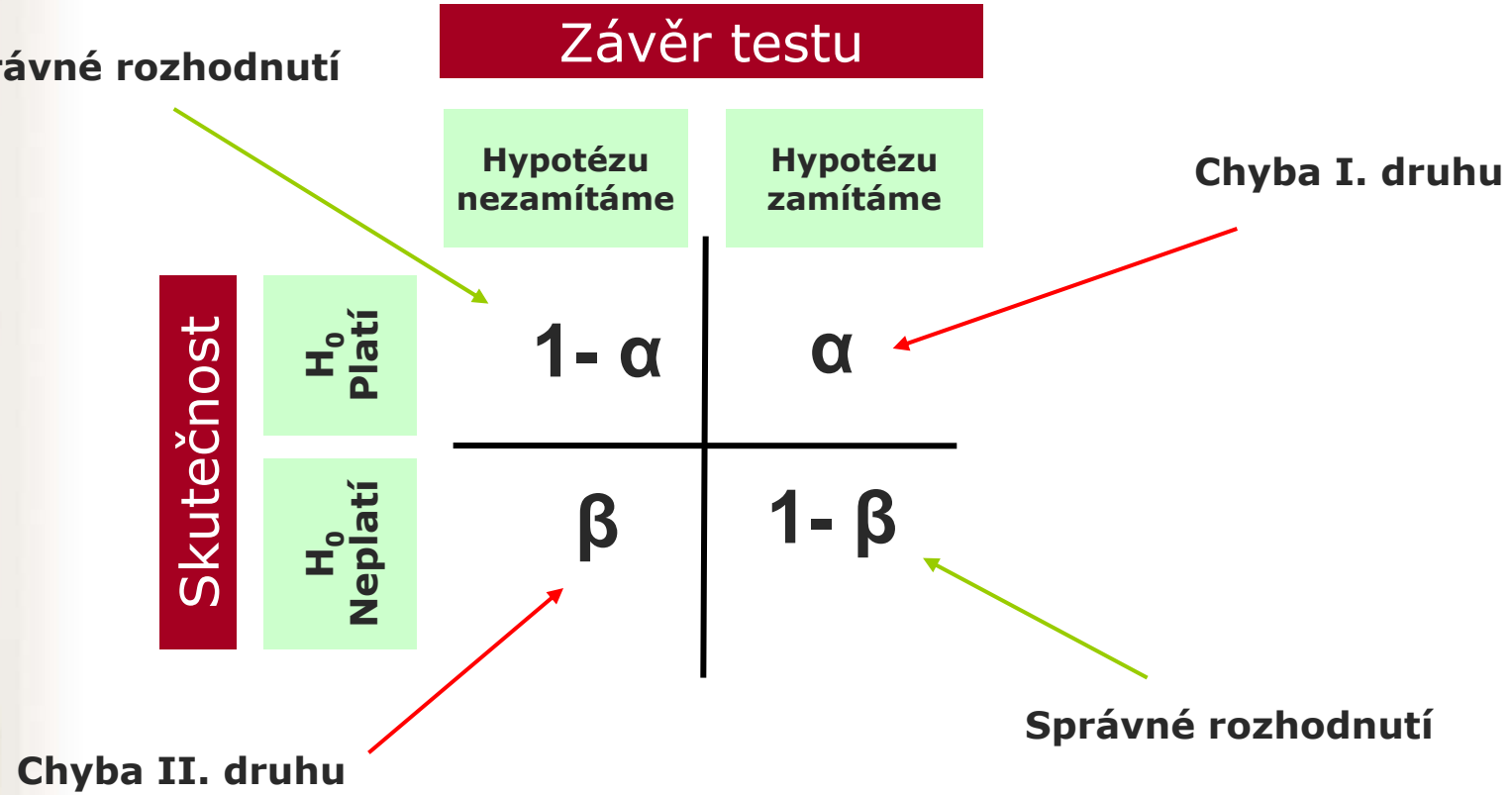
Statistické testování odpovídá na otázku zda je pozorovaný rozdíl náhodný či nikoliv. K odpovědi na otázku je využit statistický model – testová statistika.

Co znamená náhodný rozdíl?



Možné chyby při testování hypotéz

- I přes dostatečnou velikost vzorku a kvalitní design experimentu se můžeme při rozhodnutí o zamítnutí/nezamítnutí nulové hypotézy dopustit chyby.



Význam chyb při testování hypotéz

Pravděpodobnost chyby 1. druhu

α → **Pravděpodobnost nesprávného zamítnutí nulové hypotézy**

Pravděpodobnost chyby 2. druhu

β → **Pravděpodobnost nerozpoznání neplatné nulové hypotézy**

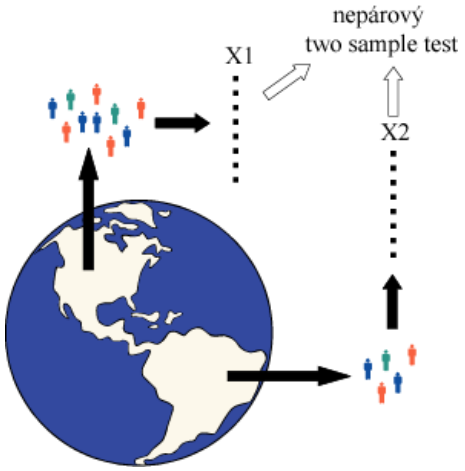
Síla testu

$1-\beta$ → **Pravděpodobnostně vyjádřená schopnost rozpoznat neplatnost hypotézy**

Nepárový vs. párový design

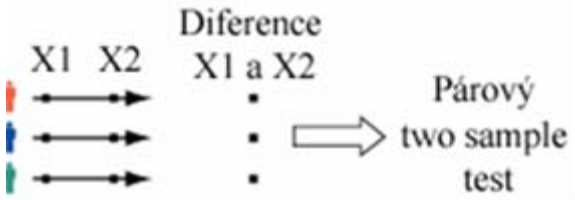
Nepárový design

- Skupiny srovnávaných dat jsou na sobě zcela nezávislé (též nezávislý, independent design), např. lidé z různých zemí, nezávislé skupiny pacientů s odlišnou léčbou atd.
- Při výpočtu je nezbytné brát v úvahu charakteristiky obou skupin dat



Párový design

- Mezi objekty v srovnávaných skupinách existuje vazba, daná např. člověkem před a po operaci, reakce stejného kmene krys atd.
- Vazba může být buď přímo dána nebo pouze předpokládána (v tom případě je nutné ji ověřit)
- Test je v podstatě prováděn na diferencích skupin, nikoliv na jejich původních datech



Statistické testy a normalita dat

- Normalita dat je jedním z předpokladů tzv. parametrických testů (testů založených na předpokladu nějakého rozložení) – např. ***t*-testy**
- Pokud data nejsou normální, neodpovídají ani modelovému rozložení, které je použito pro výpočet (*t*-rozložení) a test tak může lhát

- Řešením je tedy:
  - **Transformace dat** za účelem dosažení normality jejich rozložení
  - **Neparametrické testy** – tyto testy nemají žádné předpoklady o rozložení dat

Typ srovnání	Parametrický test	Neparametrický test
2 skupiny dat nepárově:	**Nepárový t-test**	**Mann Whitney test**
2 skupiny dat párově:	**Párový t-test**	**Wilcoxon test, sign test**
Více skupin nepárově:	**ANOVA**	**Kruskal- Wallis test**
Korelace:	**Pearsonův koeficient**	**Spearmanův koeficient**

Jednovýběrový *t*-test: příklad

- Příklad: **Nový lék na rakovinu plic** (předpokládáme studii s dostatečně velkým n)

Průměrná doba přežití pacientů je 27 měsíců

Průměrná doba přežití bez léku je 22 měsíců

prodlužuje nový lék přežití?

H_0: μ = 22,2 měsíce

H_1: μ > 22,2 měsíce

Testová statistika: **T = 6,120**

5% kritická hodnota normálního rozdělení ⟹ 1,645

Jelikož hodnota statistiky T překračuje kritickou hodnotu

Zamítáme H_0

Doba přežití léčených pacientů se oproti neléčeným prodlouží.

Výzvy statistické analýzy dat

BIOSTATISTIKA: otevřená oblast

Věda přinášející novou kvalitu

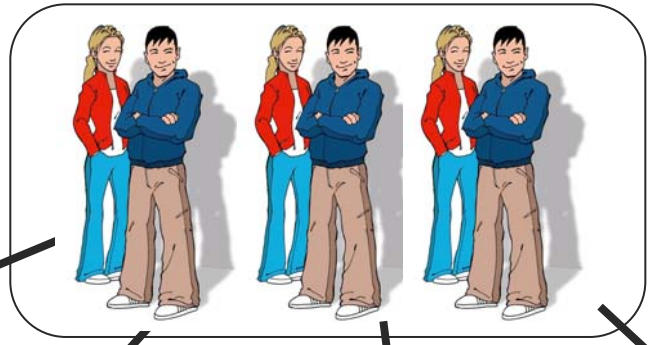
- Popisná analýza dat („exploratorní“ analýzy)
- Data mining („investigativní“ analýzy)

- Srovnávací analýzy, testy hypotéz
- Experimentální plány („experimental design“)
- QA/QC

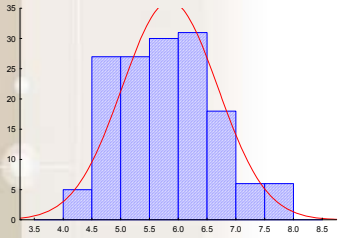
- Stochastické modelování, hodnocení prognóz
- Vícerozměrné analýzy, „pattern recognition“

- Analýza biodiverzity (species community associations, ….)
- Analýza časových řad, analýzy trendů
- Analýza biomedicínských dat

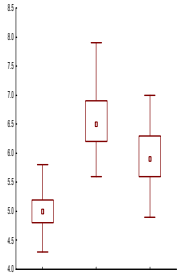
Složitý biologický systém



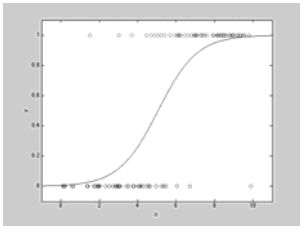
Jednorozměrná popisná statistika



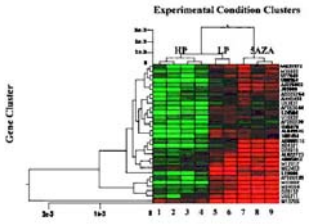
Jednorozměrné testování



Modelování



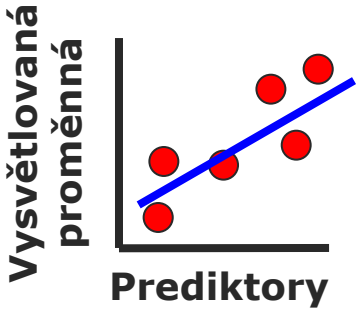
Vícerozměrná analýza a modelování



Metaanalýza

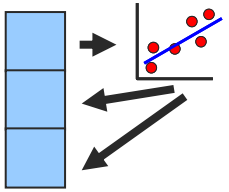
Modelování dat

1.Tvorba modelu



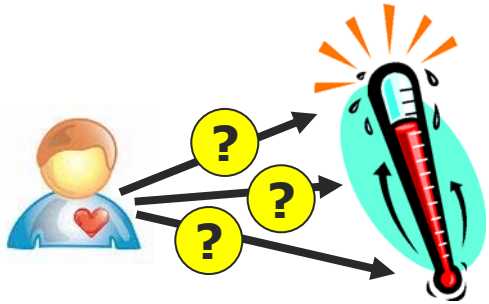
- Parametry ovlivňující vysvětlovanou charakteristiku pacienta
- Rovnice umožňující predikci
- Platnost modelu pouze v rozsahu prediktorů

2.Validace modelu



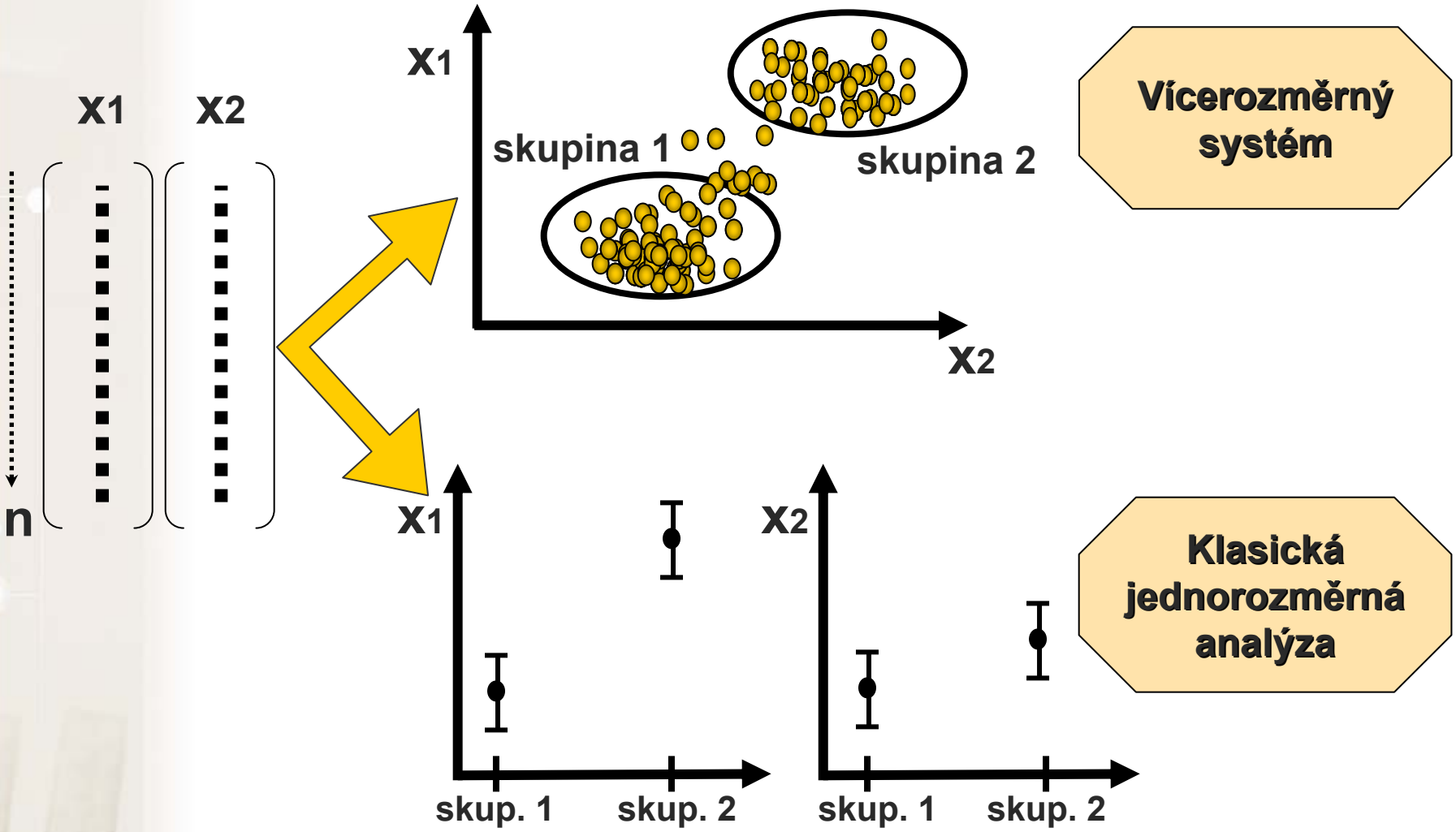
- Nebezpečí „přeučení" modelu
- Testování modelu na známých datech
- Krosvalidace

3. Aplikace modelu

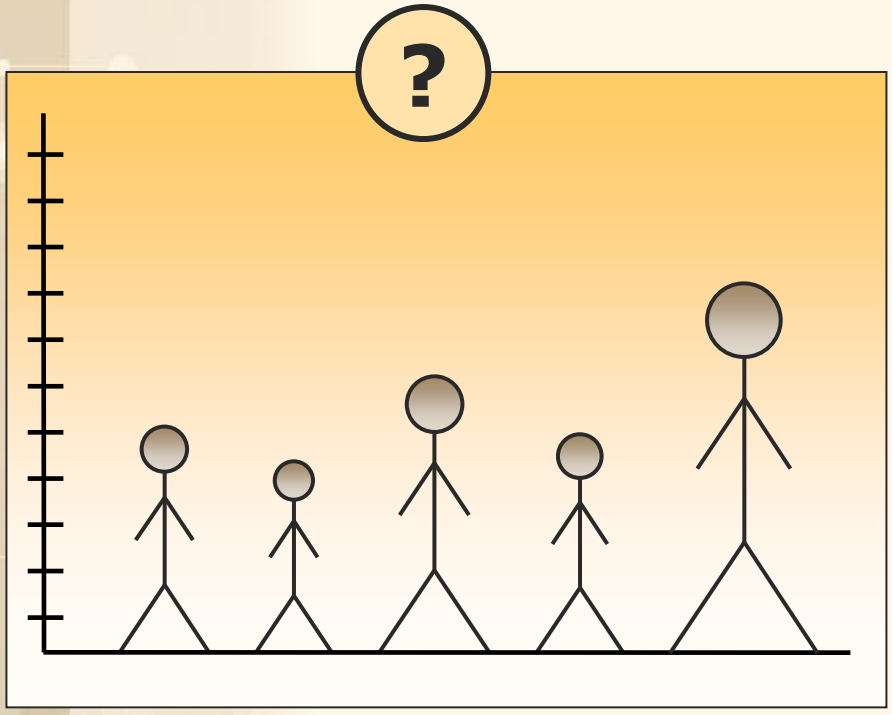


- Individuální predikce stavu neznámých případů
- Model musí být podložen korektní statistikou a rozsáhlými daty

Vícerozměrné vnímání skutečnosti – nová kvalita analýzy dat



Běžná sumarizace dat „likviduje“ individualitu jedince



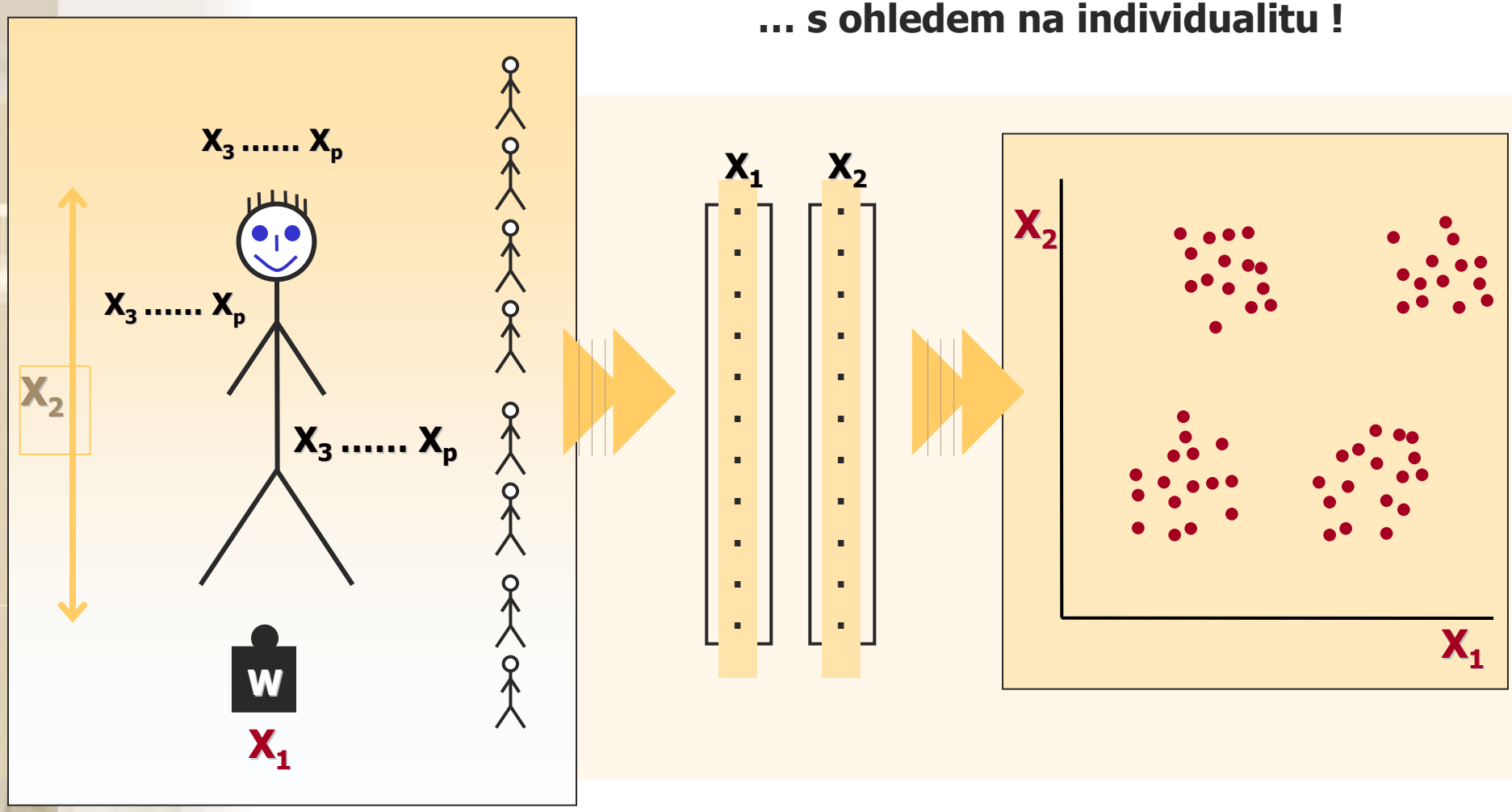
Průměr ± SE

BĚŽNÁ STATISTICKÁ SUMARIZACE

- ✓ *Zpřehlednění dat*
- ✓ *Neodliší původní měření*

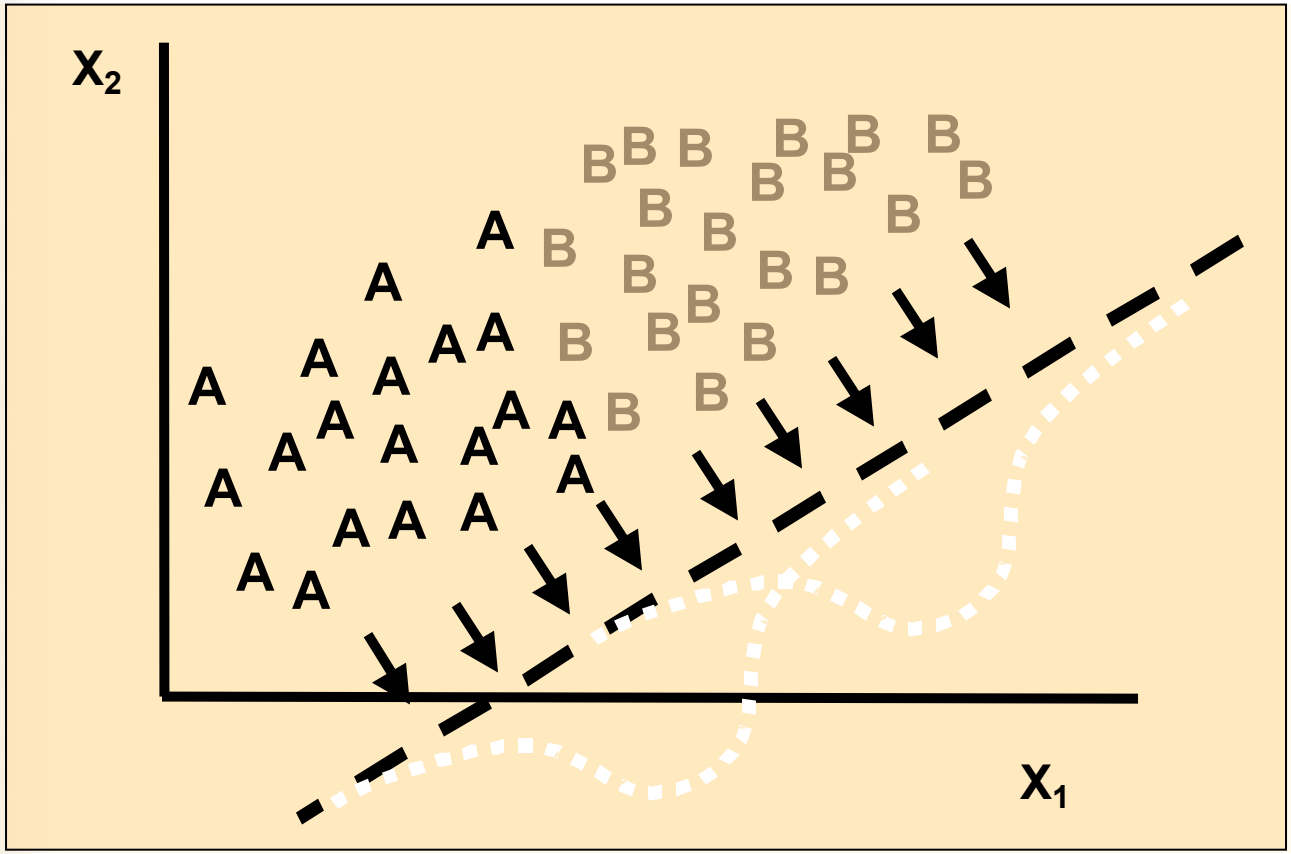
Vícerozměrné hodnocení

… s ohledem na individualitu !



Vícerozměrné hodnocení – nová kvalita

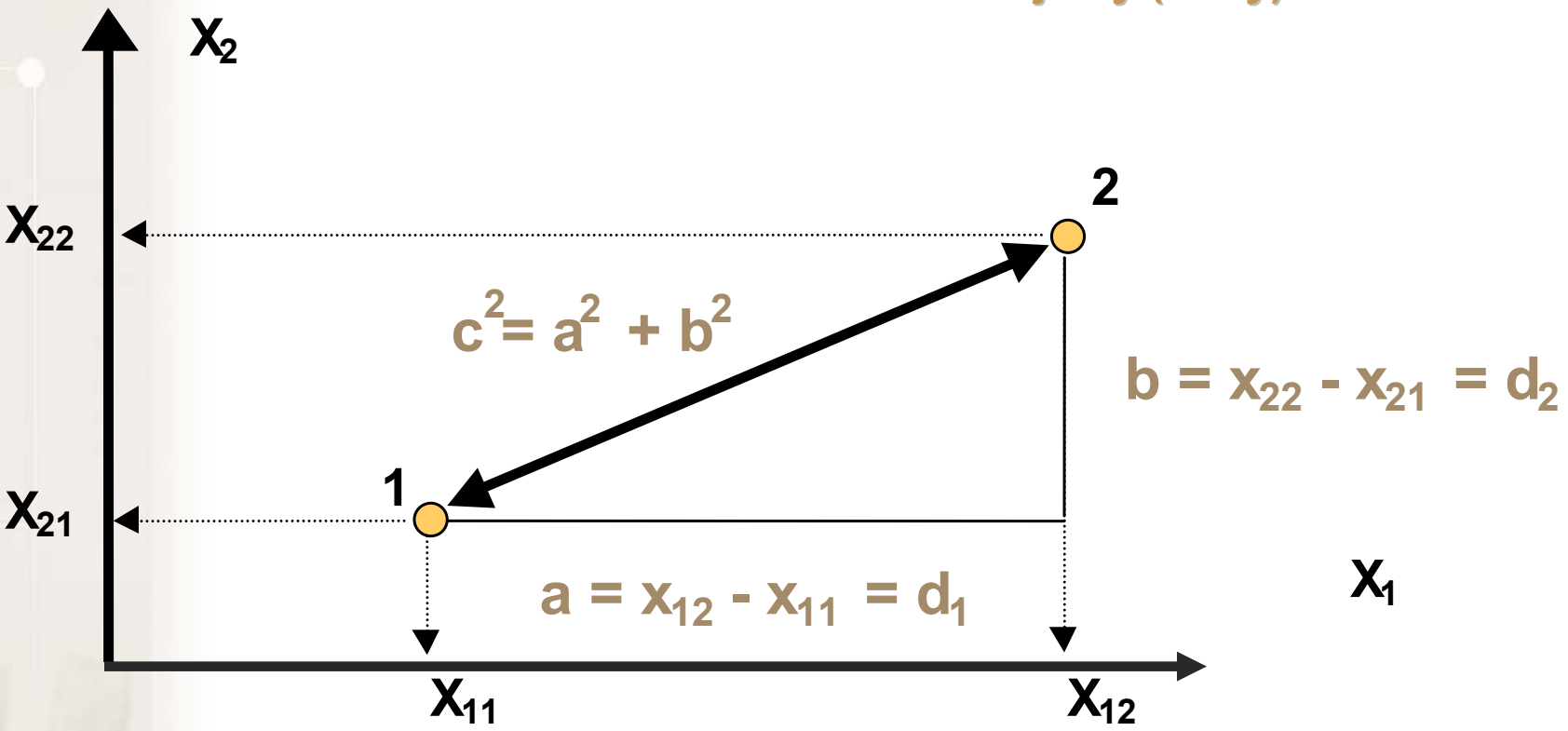
Pouze kombinované parametry mají odpovídající informační sílu



příklad: X1 =

Vícerozměrné hodnocení vychází z jednoduchých principů

příklad: vícerozměrná vzdálenost měření mezi dvěma objekty (body)



Vícerozměrné modelování je strategickou disciplínou

$X_1 \ldots\ldots X_n$

technické parametry automobilu

$X_{n+1} \ldots\ldots X_p$

řidičovy schopnosti a jeho stav

$X_{p+1} \ldots\ldots X_2$

rychlost, povrch, situace

$X_1 \quad X_2 \quad X_3 \quad X_4 \quad X_5 \quad \ldots\ldots \quad X_p$

Základní typy vícerozměrných analýz

SHLUKOVÁ ANALÝZA

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

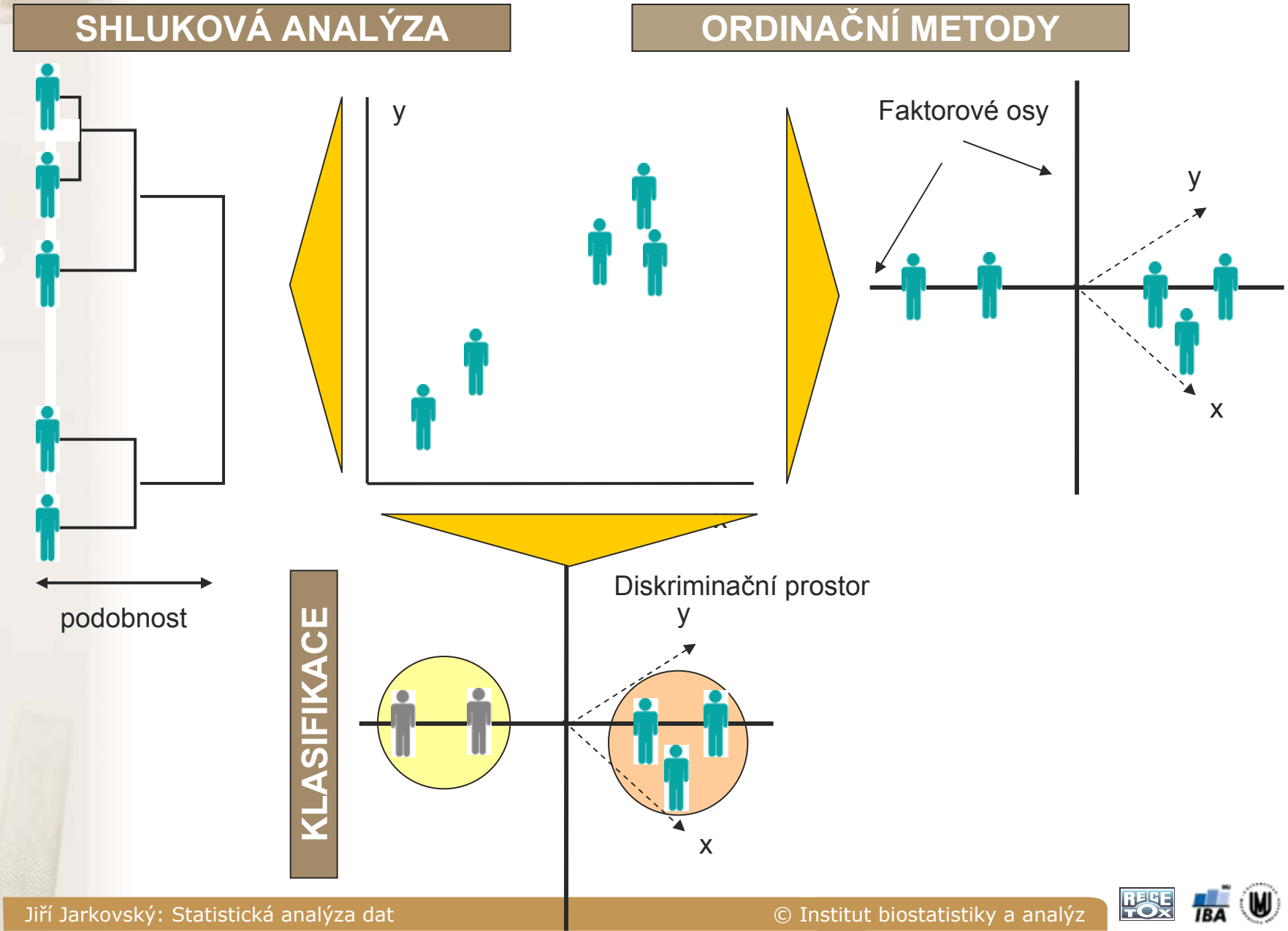
KLASIFIKACE

- Model zařazení neznámých pacientů do předem daných skupin
- Řada algoritmů

ORDINAČNÍ METODY

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

Typy vícerozměrných analýz



Shrnutí

Software pro statistickou analýzu

- Komplexní software pro všechny úkoly neexistuje
- Obecné komerční statistické balíky jako je **SPSS, SAS, Statistica, S+, ArcGis** nabízí širokou škálu metod v uživatelsky přívětivém prostředí,ale … **některé specializované metody nejsou k dispozici**
- Specializované nástroje (freeware i komerční) jsou nezbytné pro některé analýzy– **R (www.r-project.org)** je dobrým řešením pro specializované analýzy

Standardní statistické balíky pro rutinní úlohy

Otevřený modifikovatelný SW pro specializované analýzy

Potřebné nástroje jsou dostupné

Shrnutí

- ☑ **Statistická analýza je nezbytná na všech úrovních výzkumu**
- ☑ **Statistická analýza je pouhým nástrojem, který má za úkol data zpřehlednit, zviditelnit a sumarizovat. Požadovány jsou nezkreslené, reprezentativní a spolehlivé závěry.**
- ☑ **Budoucnost je v individuálním posuzování vývoje jednotlivých případů – vícerozměrná analýza a modelování.**

Enjoy your analyses !!!