Biostatistika

jarkovsky@iba.muni.cz

Přednáška 1

Organizační informace – výukové materiály

- Tato prezentace v IS.MUNI + prezentace a příklady ovládání SW Statistica + další souhrnné podklady
- www.matematickabiologie.cz/res/file/ucebnice/pavlik-biostatistika.pdf
- portal.matematickabiologie.cz/index.php?pg=aplikovana-analyza-klinickych-a-biologickych-dat--biostatistika-pro-matematickou-biologii
- Tabulky statistických rozdělení
- Libovolná základní učebnice statistiky – např.
  - https://www.amazon.com/Biostatistical-Analysis-5th-Jerrold-Zar/dp/0131008463/ref=sr_1_1?ie=UTF8&qid=1505890489&sr=8-1&keywords=zar+biostatistical+analysis
  - https://www.amazon.com/Medical-Statistics-Glance-Aviva-Petrie/dp/140518051X/ref=sr_1_sc_1?s=books&ie=UTF8&qid=1505890508&sr=1-1-spell&keywords=avive+petria
  - https://www.amazon.com/Statistics-Veterinary-Animal-Science-Petrie/dp/0470670754/ref=sr_1_sc_3?s=books&ie=UTF8&qid=1505890522&sr=1-3-spell&keywords=avive+petria

Organizační informace – software

- Software
  - Univerzitní licence na inet.muni.cz (stejný login a passwd jako do is.muni.cz)
  - Statistica – www.statsoft.com, www.statsoft.cz
  - SPSS - www.ibm.com/analytics/us/en/technology/spss/
  - R – www.r-project.org, www.rstudio.com
  - Stata - www.stata.com

Statistika ve vědecké praxi

Pozice statistické analýzy ve vědě a klinické praxi

Význam statistických výstupů

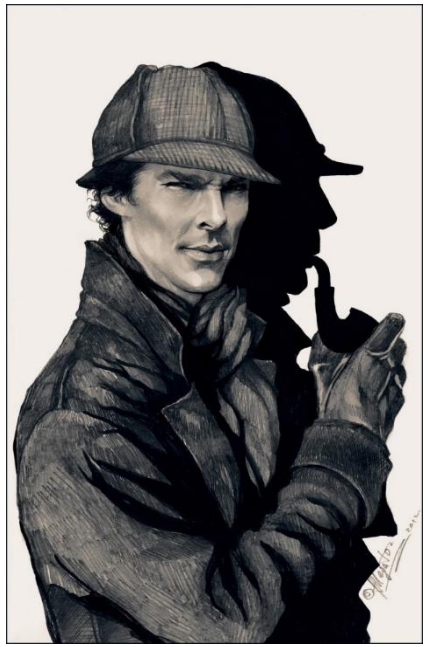
Anotace

- Statistická analýza biologických dat je jedním z nástrojů, s jejichž pomocí se snažíme zjistit odpovědi na naše otázky týkající se pochopení živé přírody.
- Jako každý nástroj je i statistickou analýzu nezbytné na jedné straně korektně využívat a na druhou stranu nepřeceňovat její možnosti.
- Klíčovým faktem při statistické analýze dat je nahlížení na realitu prostřednictvím vzorku a přijmutí toho, že výsledky naší analýzy jsou jen tak dobré, jak dobrý je náš vzorek.
- Reprezentativnost, nezávislost a náhodnost vzorku spolu s jeho velikostí jsou důležité faktory ovlivňující věrohodnost našich závěrů.

Life is beautiful with data analysis

Data jsou základ vědecké práce

- *Data! Data! Data! I can't make bricks without clay!*
  Sir Arthur Conan Doyle

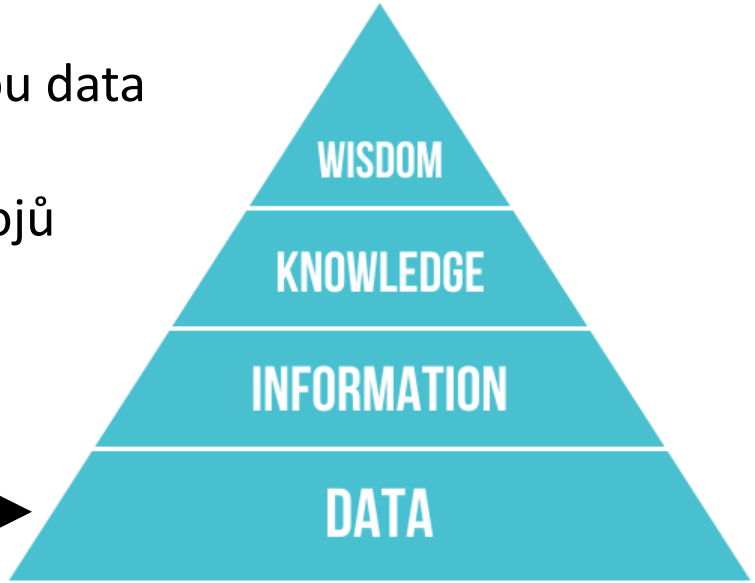


Základem pro naši práci jsou data

Získáváme je z různých zdrojů

A musíme s nimy neustále pracovat

DIKW pyramid



Co znamená pro biologa/lékaře statistická analýza dat?

- **Matematická statistika** je vědecká disciplína na pomezí popisné statistiky a aplikované matematiky. Zabývá se teoretickým rozborem a návrhem metod získávání s analýzy empirických dat obsahujících prvek nahodilosti, tedy teorií plánování experimentů, výběrů, statistických odhadů, testování hypotéz a statistických modelů.
- **Statistika** je věda a postup jak rozvíjet lidské znalosti použitím empirických dat. Je založena na matematické statistice, která je větví aplikované matematiky.
- **Biostatistika** = aplikace statistické analýzy dat v biologickém a klinickém výzkumu
  - Nástroj pro uchopení dat našeho výzkumu
  - Nezbytné chápat principy a limitace
  - Není nutná detailní matematická znalost

- **Easy to understand, hard to master**

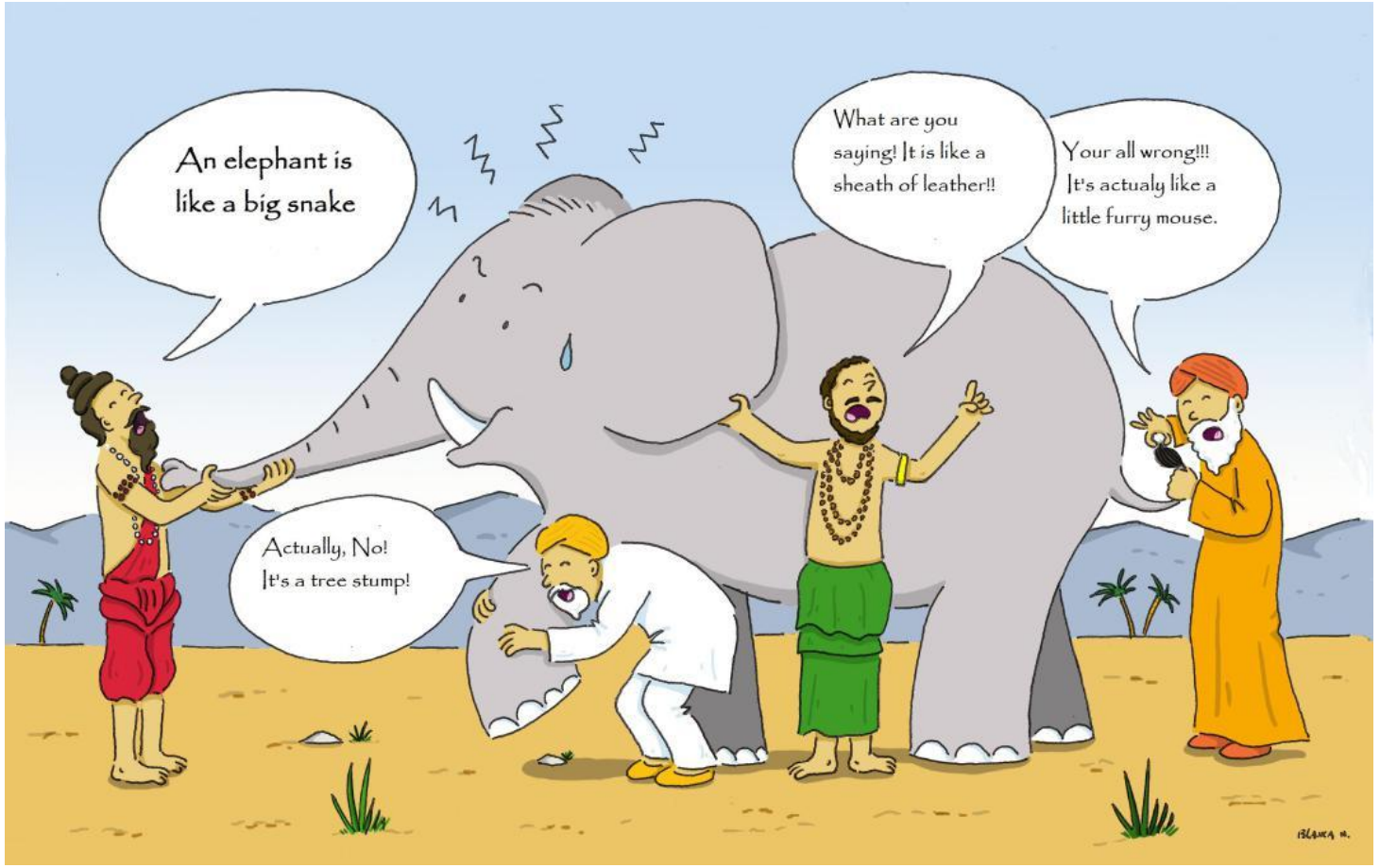
Výzkum, realita, statistika

- Výzkum je naším způsobem porozumění realitě
- Ale jak přesné a pravdivé je naše porozumění?

- **Statistika** je jedním z nástrojů umožňujícím popis a komunikaci výsledků výzkumu.
- Ale je to pouze nástroj, co je skutečně důležité jsou **data**.

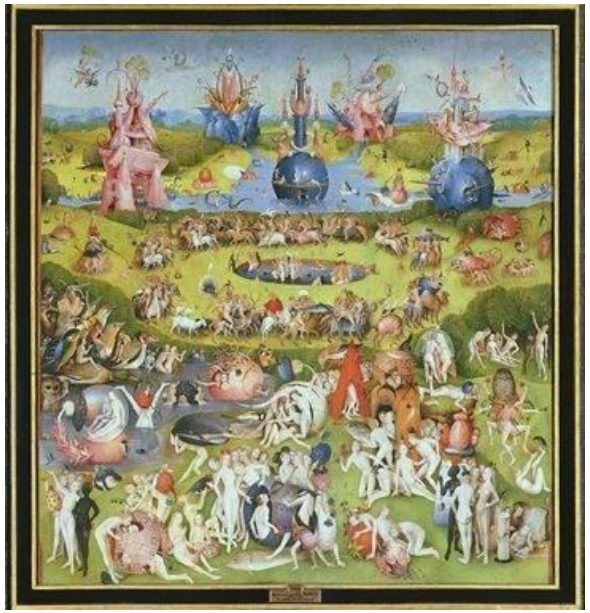
Realita a data

- Klíčovou otázkou výzkumu a následně statistické analýzy je jak dobře naše data popisují realitu
- Bez kvalitních dat není kvalitní statistiky ani kvalitního výzkumu.
- Každá chyba učiněná v úvodní fázi výzkumu se v dalších fázích znásobí a zřejmě ji již nebude možné eliminovat



Variabilita jako základní pojem ve statistice

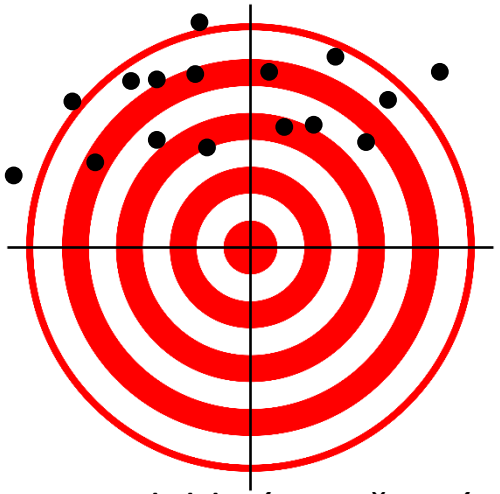
- Naše realita je variabilní a statistika je vědou zabývající se variabilitou
- Korektní analýza variabilita a její pochopení přináší užitečné informace o naší realitě
- V případě deterministického světa by statistická analýza nebyla potřebná



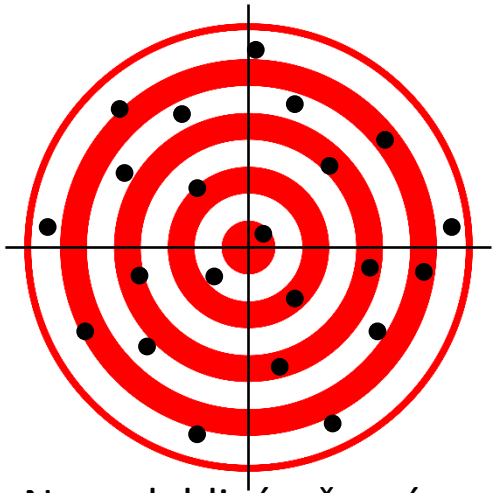
?

Spolehlivost a přesnost měření

- Kvalita dat je klíčová pro jakékoliv statistické hodnocení
- Bez spolehlivých a přesných dat není možné získat spolehlivé a přesné výsledky statistického hodnocení
- Ve statistické analýze dat musíme zohlednit jak střed měření, tak variabilitu a zamyslet se nad přesností popisu reality

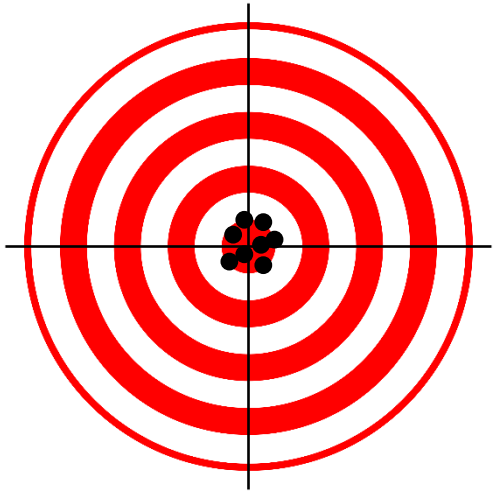


Nespolehlivý, nepřesný



Nespolehlivý, přesný

Spolehlivý, nepřesný



Spolehlivý, přesný

Variabilita a střední hodnota

- Norma = 5 gramů soli na 1 kg rýže

Nezamícháte

0g soli / 1 kg rýže

10g soli / 1 kg rýže

Průměr: 5g soli / 1 kg rýže

Vše OK !!!

Průměr není vše, je nezbytné zohlednit variabilitu

Zamícháte

5g soli / 1 kg rýže

5g soli / 1 kg rýže

Průměr: 5g soli / 1 kg rýže

Vše OK !!!

Nárůst šumu s detailem

Častý požadavek na stále detailnější a detailnější výstupy vede k nesmyslným a zavádějícím výsledkům.

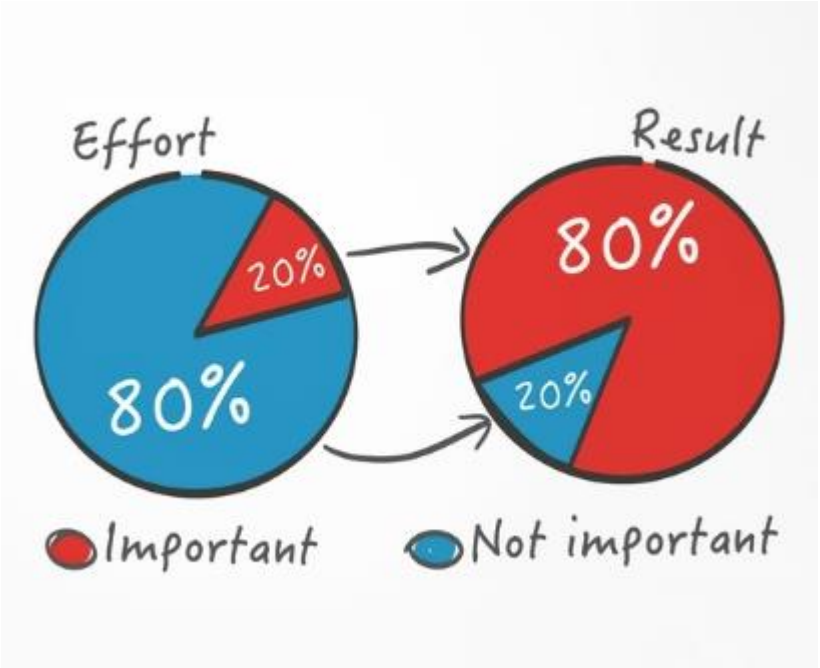
Např. proočkovanost nad 100% obyvatel – zní podezřele, ale je důsledkem nedostatečné přesnosti demografických dat na úrovni malých obcí.

Paretovo pravidlo v praxi

V reálu sbíraná data obsahují vždy nějaký šum

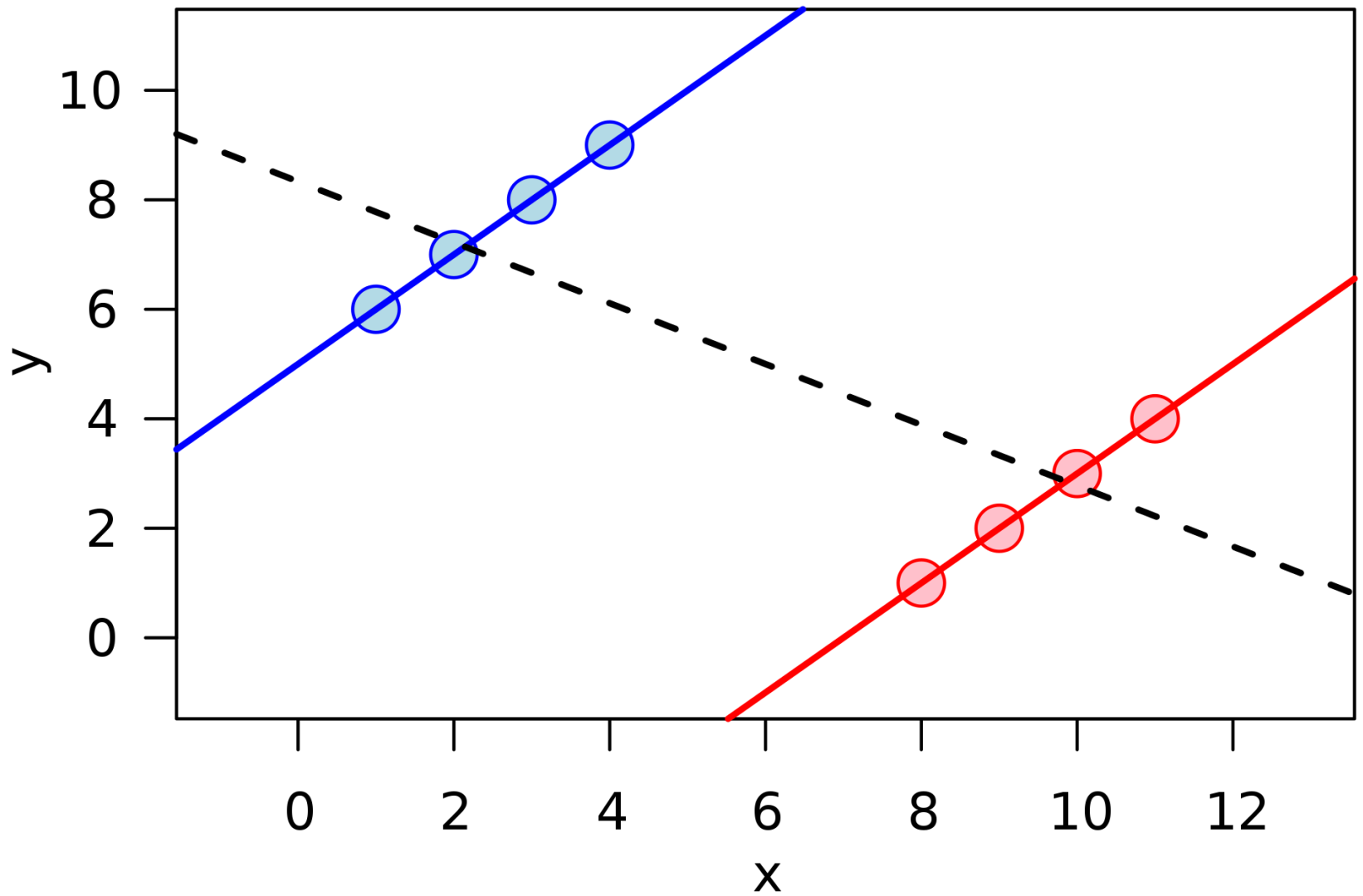
Projeví se zejména při velmi detailním pohledu

Odstranění veškerého šumu je velmi časově náročné a v praxi v podstatě neproveditelné z pohledu dostupných kapacit



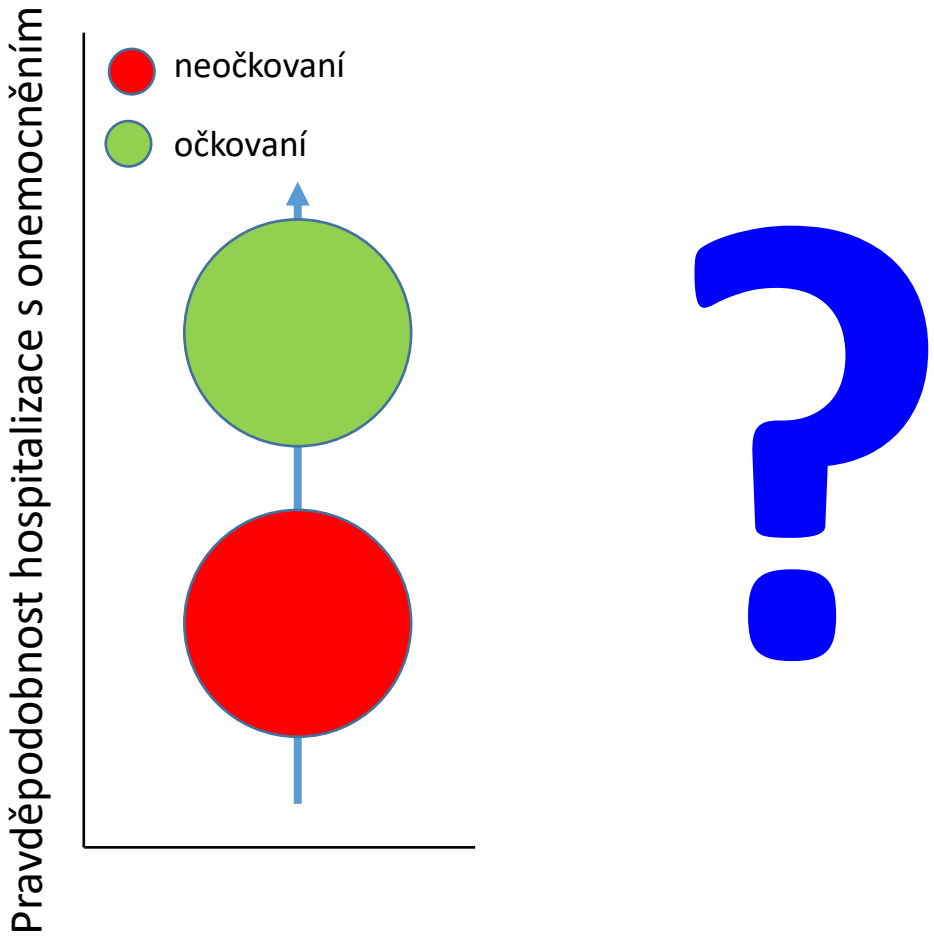
Je třeba si být vědom nedokonalostí dat v detailech.

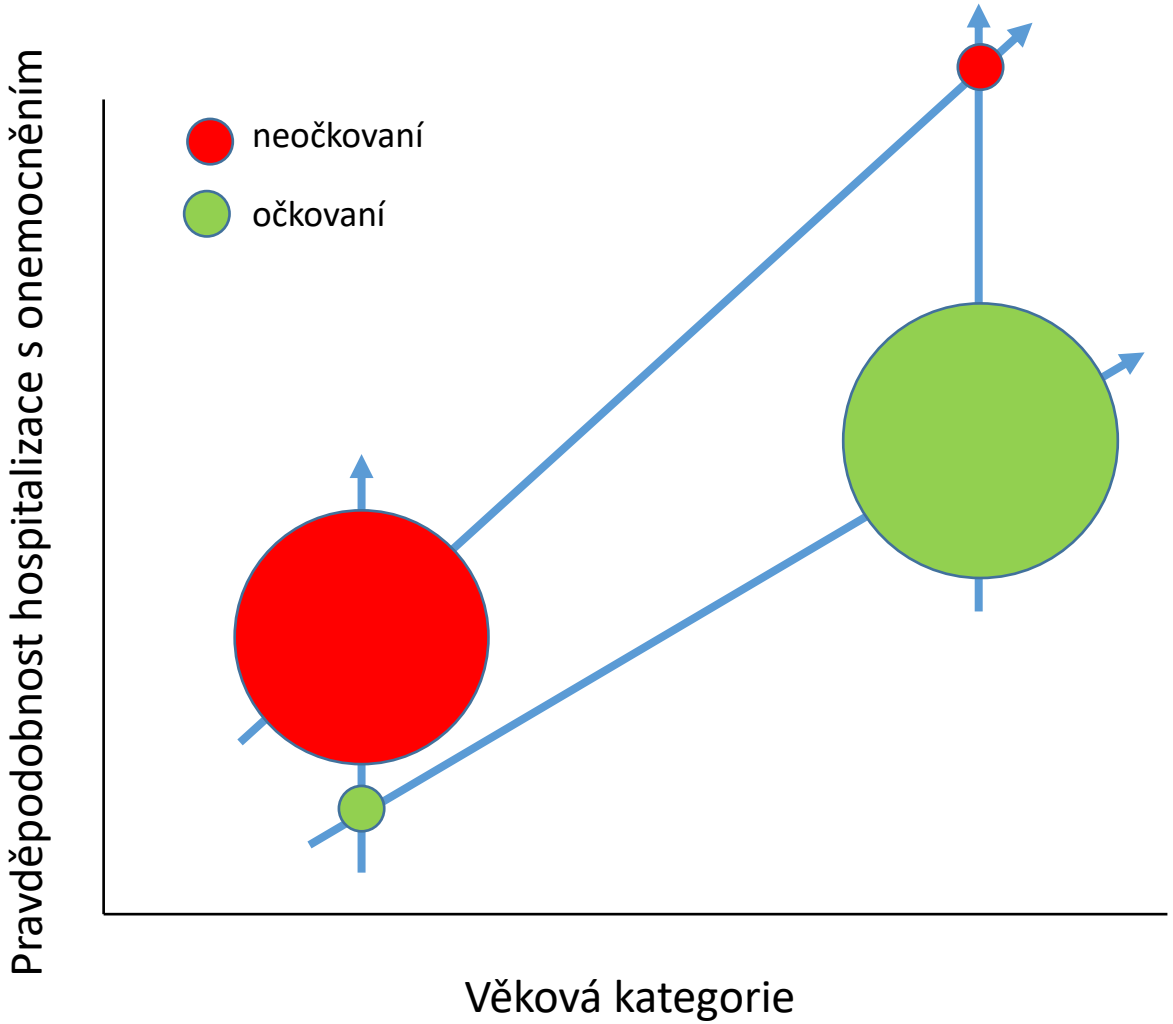
Interpretujeme výsledky správně? Simpson paradox



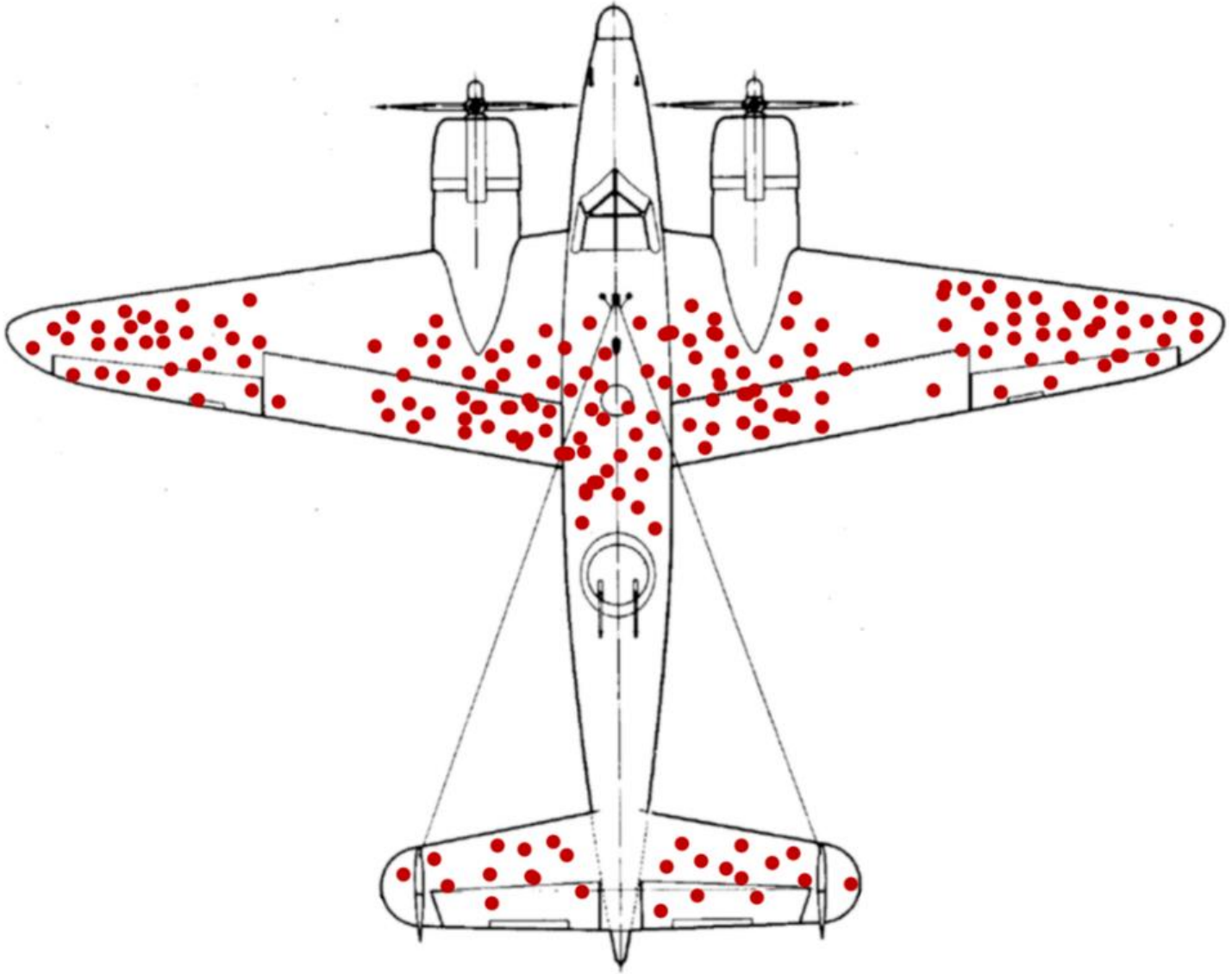
Simpson paradox

Typický problém chybné interpretace dat, velmi snadno vzniká pokud nedošlo k pochopení podstaty dat.



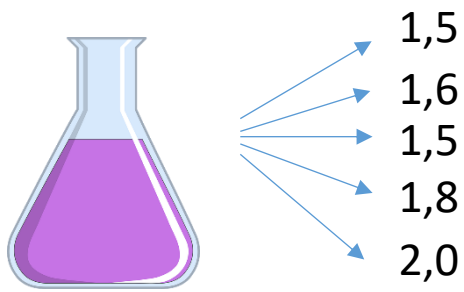


Je realita skutečně realita? Survivor bias

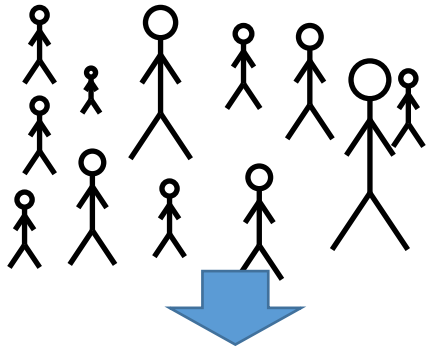


Různé úrovně variability

Variabilita opakovaných měření

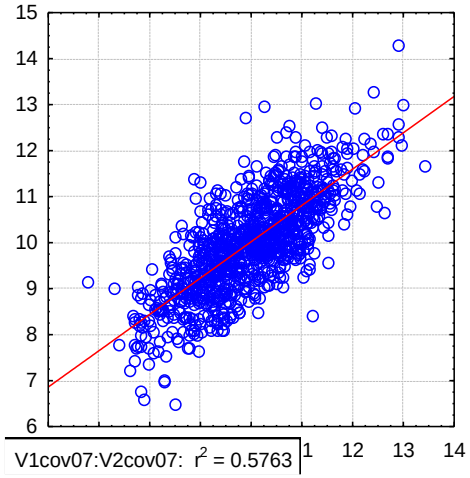


Variabilita dat v populaci

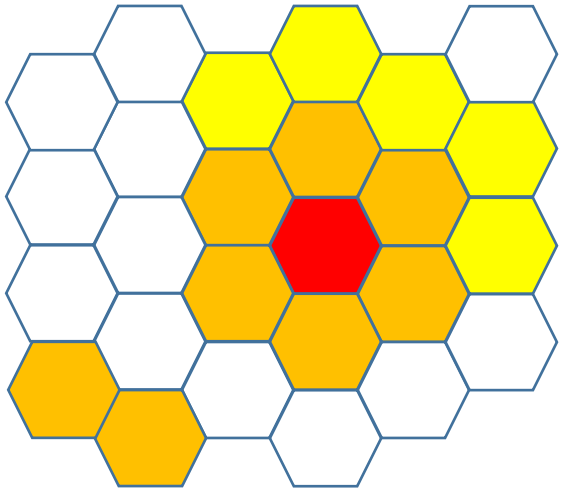


Hlavní téma kurzu

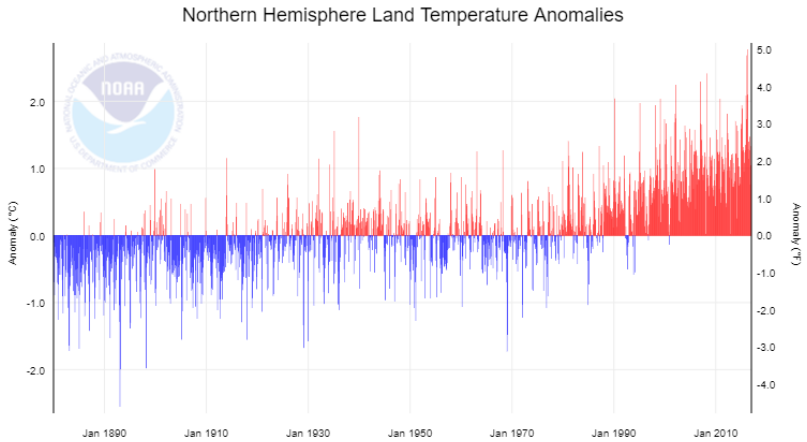
Variabilita v modelech



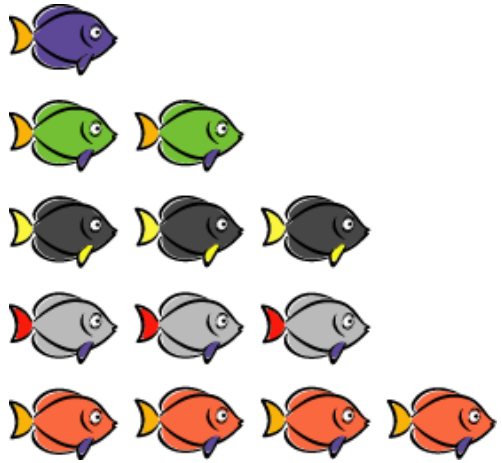
Geografická variabilita



Variabilita časových řad

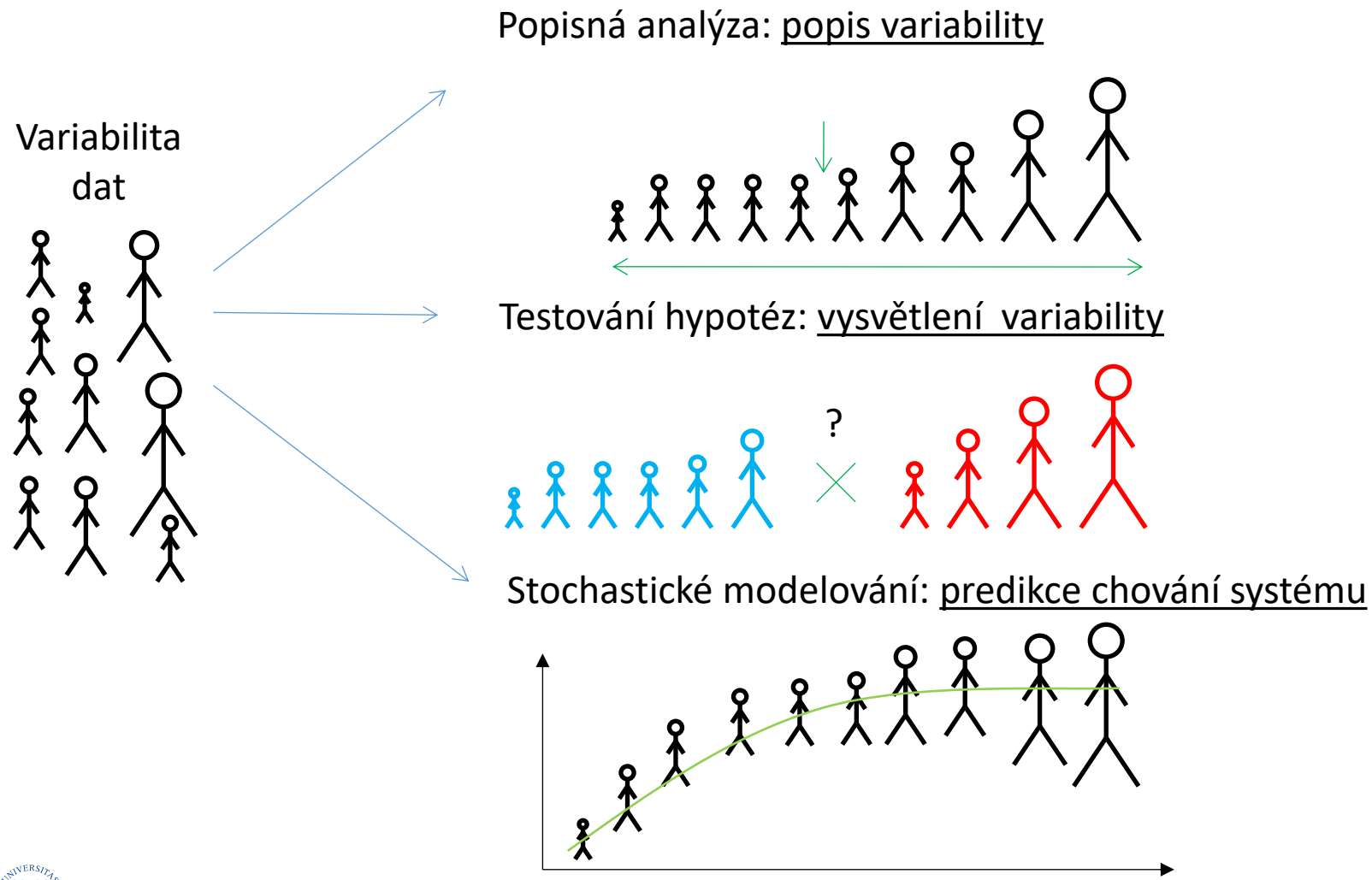


Biodiverzita



Práce s variabilitou v analýze dat

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



Statistika – definice

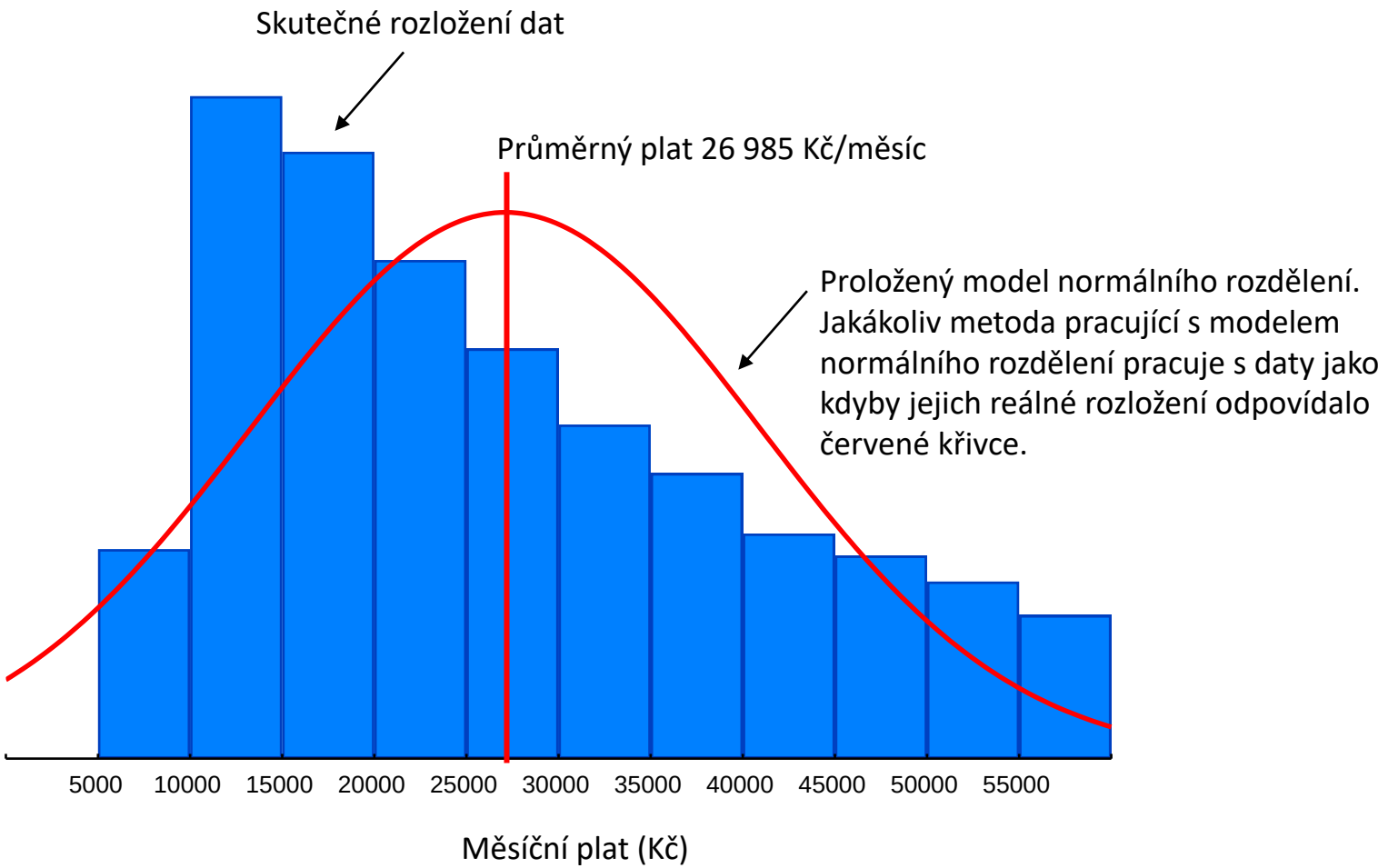
WWW.WIKIPEDIA.ORG:

Statistika je matematickou vědou zabývající se shromážděním, analýzou, interpretací, vysvětlením a prezentací dat. Může být aplikována v širokém spektru vědeckých disciplín od přírodních až po sociální vědy. Statistika je využívána i jako podklad pro rozhodování, kdy nicméně může být záměrně i nevědomky zneužita.

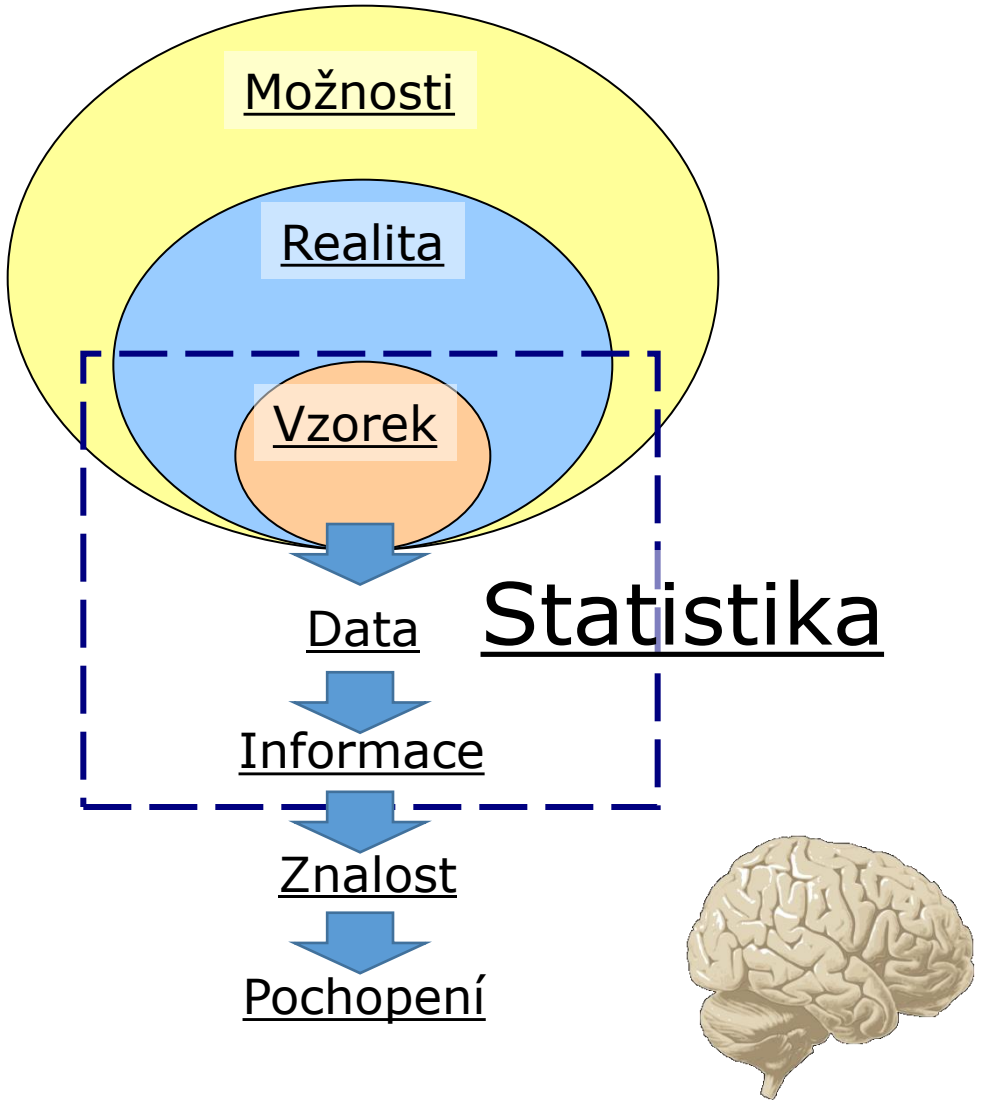
Statistika využívá matematické modely reality k zobecnění výsledků experimentů a vzorkování. Statistika funguje korektně pouze pokud jsou splněny předpoklady jejích metod a modelů.

Nesprávná aplikace modelu -> zkreslené závěry

- Různé popisné statistiky a testy jsou spjaty s různými modelovými rozděleními
- Pro správnou interpretaci je třeba ověřit shodu reálných dat s modelem
- Některé statistiky je možné vždy spočítat, ale jejich interpretace je v případě nedodržení předpokladů pouze omezená



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 !!!

Co musíme vědět před zahájením studie nebo experimentu?

- Cílová populace
  - Skupina objektů (pacientů, lokalit atd.) na něž je studie zaměřena

- Primární hypotézy
  - Hlavní otázka položená ve studii – odhad velikosti vzorku a design studie je vypracován vzhledem k primární hypotéze (v řadě případů nelze v reálném výzkumu formální power analýzu vypracovat, nicméně zamyšlení nad velikostí vzorku je nezbytné vždy)

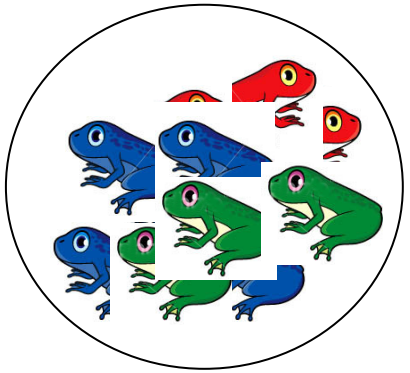
- Sekundární hypotézy
  - Vedlejší otázky, na něž by studie měla odpovědět

- Výběr adekvátní metodiky
  - Hypotézy jsou zodpovězeny prostřednictvím konkrétních proměnných (endpointů) – jejich typ (binární, kategoriální, spojité proměnné, biodiverzita, přežití, mortalita atd.) určuje výběr způsobu statistického zpracování

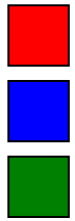
Cílová populace

- Cílová populace – klíčový pojem statistického zpracování
  - Skupina objektů o nichž se chceme něco dozvědět (např. lokality v daném povodí, laboratorní organismy v daných podmínkách, pacienti s danou diagnózou, všichni lidé nad 60 let, měření hemoglobinu v dané laboratoři)
  - Musí být definována ještě před zahájením sběru dat
  - Na cílové populaci probíhá vzorkování dat, které musí cílovou populaci dobře (reprezentativně) charakterizovat

Cílová populace



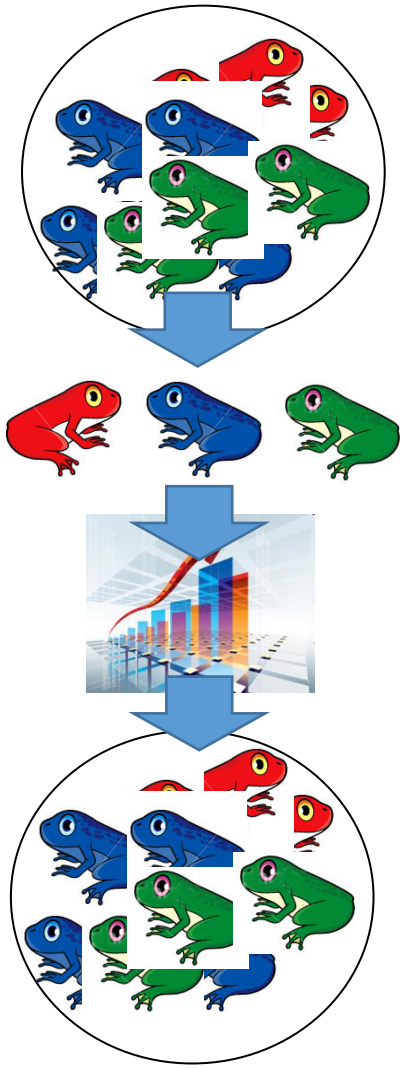
Klíčové faktory cílové populace



Design experimentu a vzorkovací plán

Vzorkování a analýza dat

Statistika a zobecnění výsledků



Neznámá cílová populace

Vzorek

Analýza

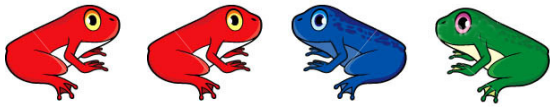
Díky zobecnění výsledků známe vlastnosti cílové populace

- Cílem analýzy není pouhý popis a analýza vzorku, ale zobecnění výsledků ze vzorku na jeho cílovou populaci
- Pokud vzorek nereprezentuje cílovou populaci, vede zobecnění k chybným závěrům

Vzorkování a jeho význam ve statistice

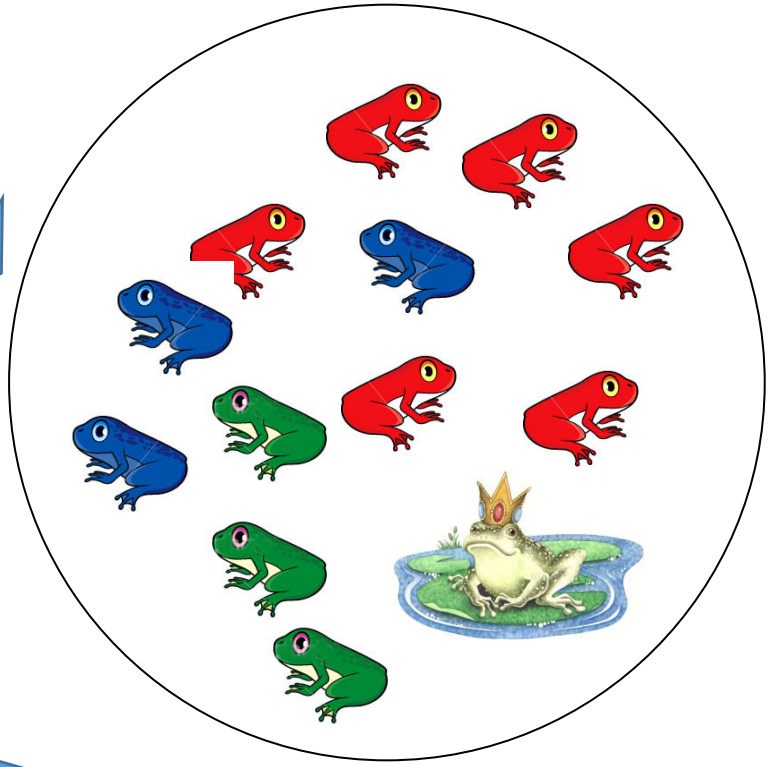
- Statistika hovoří o realitě prostřednictvím vzorku!!!
- Statistické předpoklady korektního vzorkování

- **Representativnost:** struktura vzorku musí maximálně reflektovat realitu

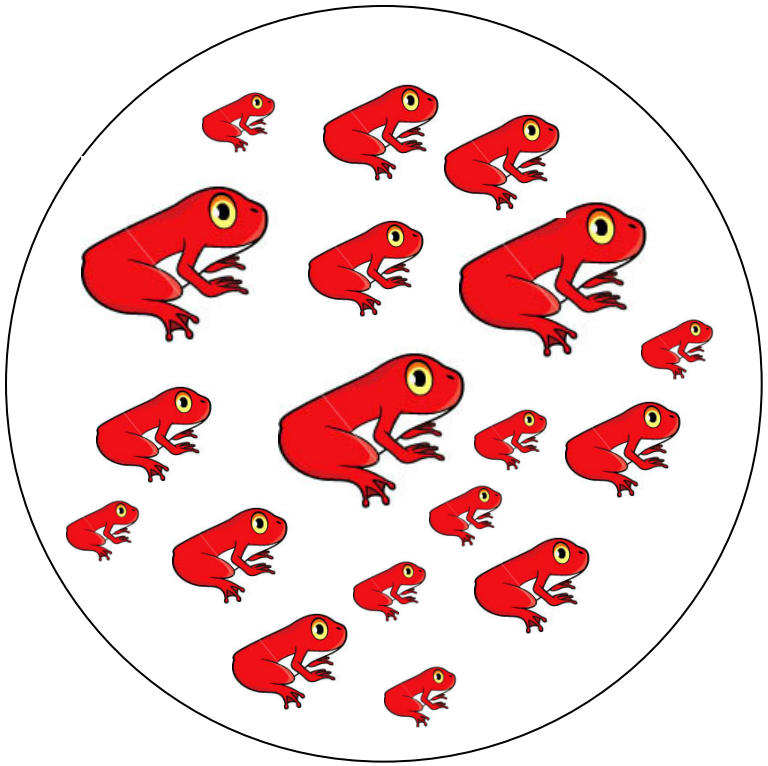


- **Nezávislost:** několikanásobné vzorkování téhož objektu nepřináší ze statistického hlediska žádnou novou informaci

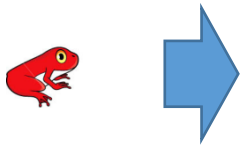
- **Náhodnost:** zajišťuje náhodný vliv zavádějících faktorů



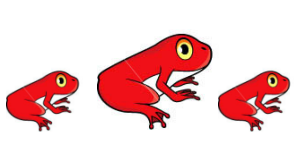
Velikost vzorku a spolehlivost statistických výstupů



- Existuje skutečné rozložení a skutečná střední hodnota měřené proměnné
- Z jednoho měření nezjistíme nic



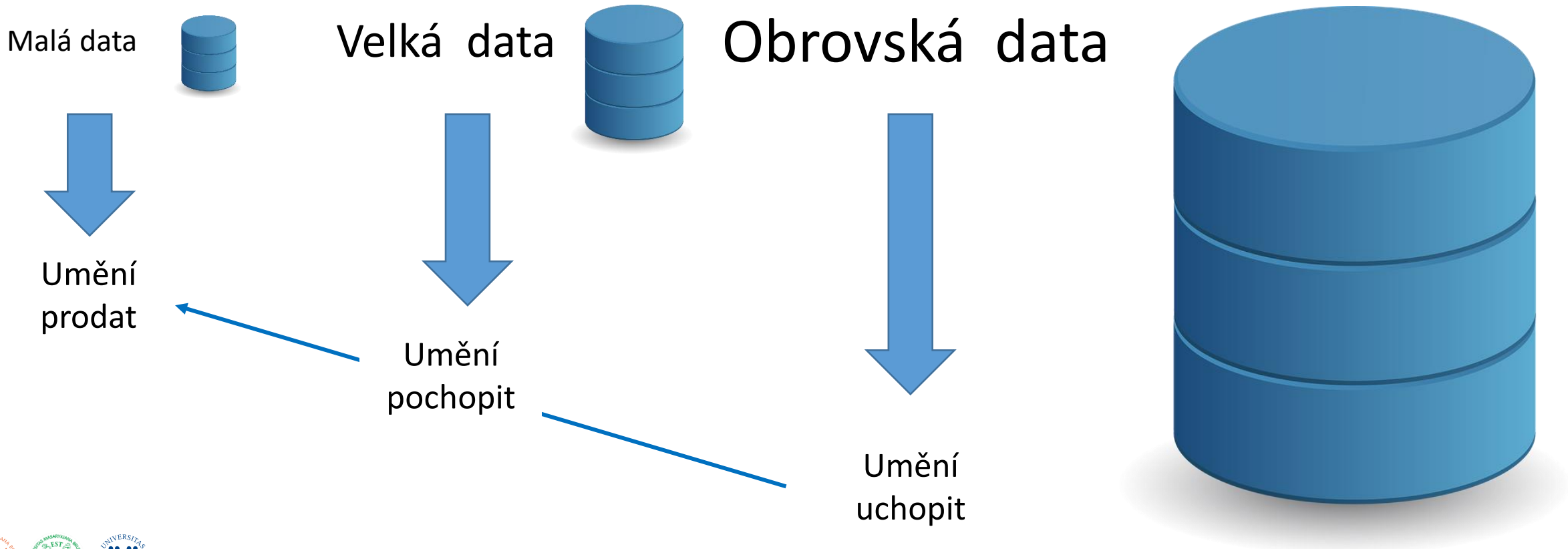
- Vzorek určité velikosti poskytuje odhad reálné hodnoty s definovanou spolehlivostí



- Vzorkování všech existujících objektů poskytne skutečnou hodnotu dané popisné statistiky, nicméně tento přístup je ve většině případech nereálný.

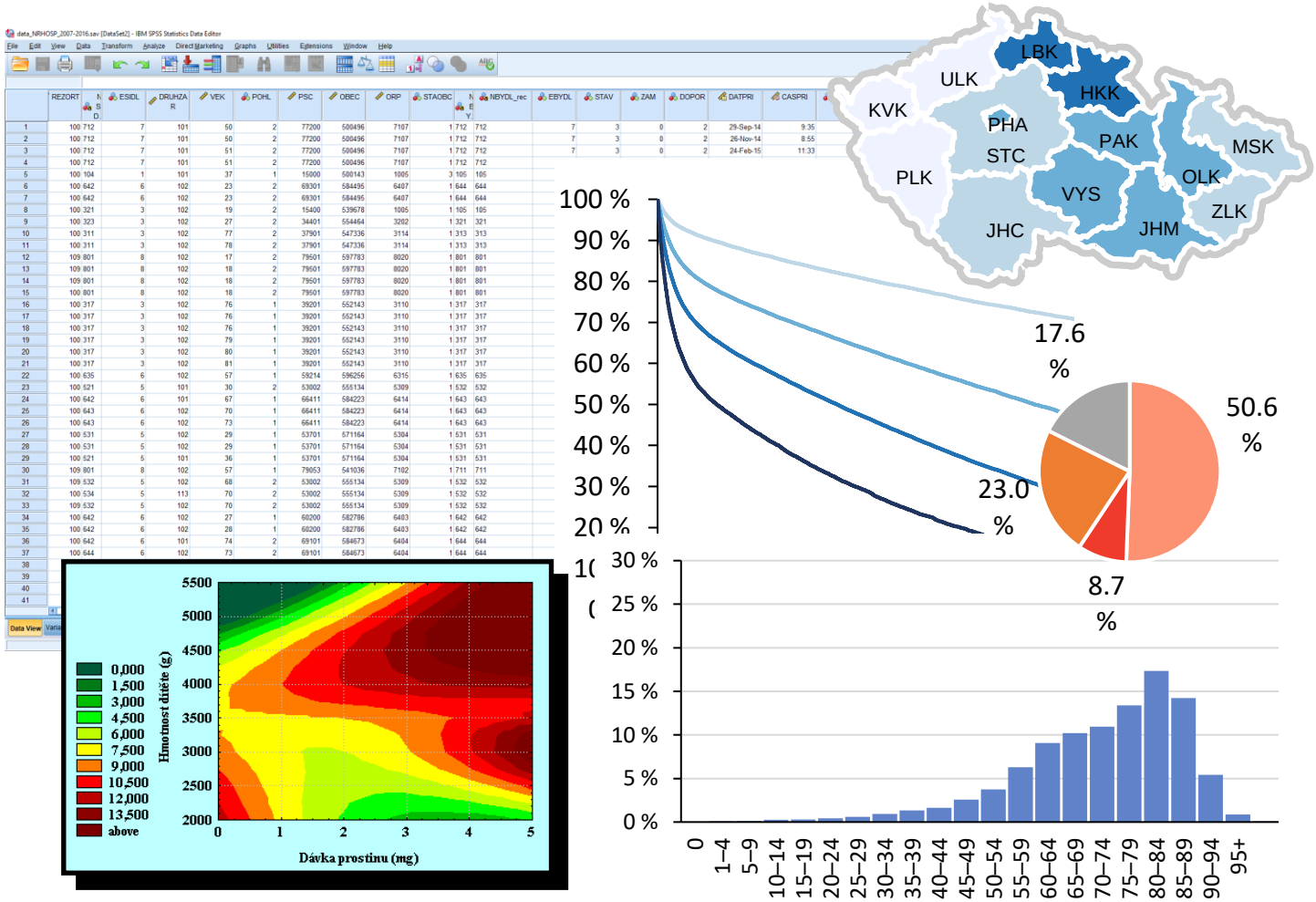
Různá velikost vzorku – různé úkoly analýzy dat

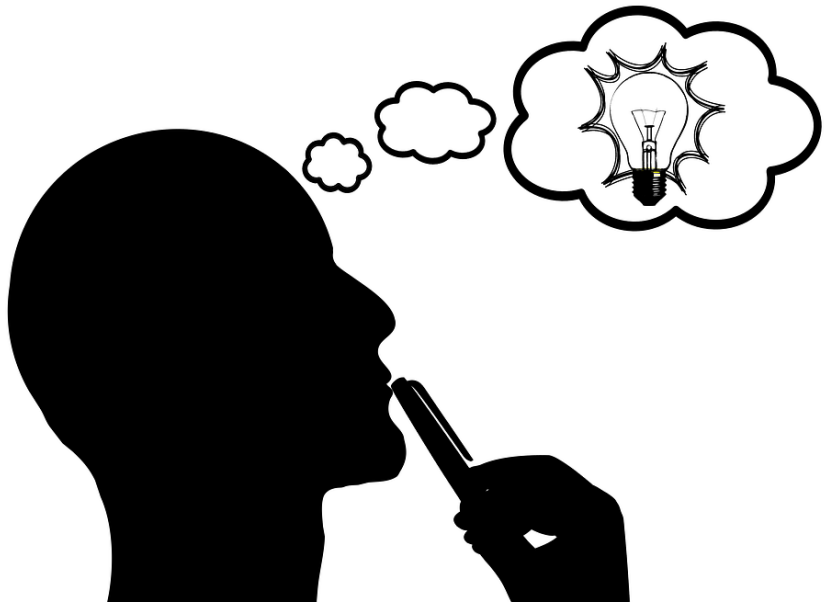
- Náročnost analýzy dat stoupá i s jejich objemem
- I u největších dat stále platí, že klíčová je schopnost data prodat = smysluplně interpretovat a prezentovat



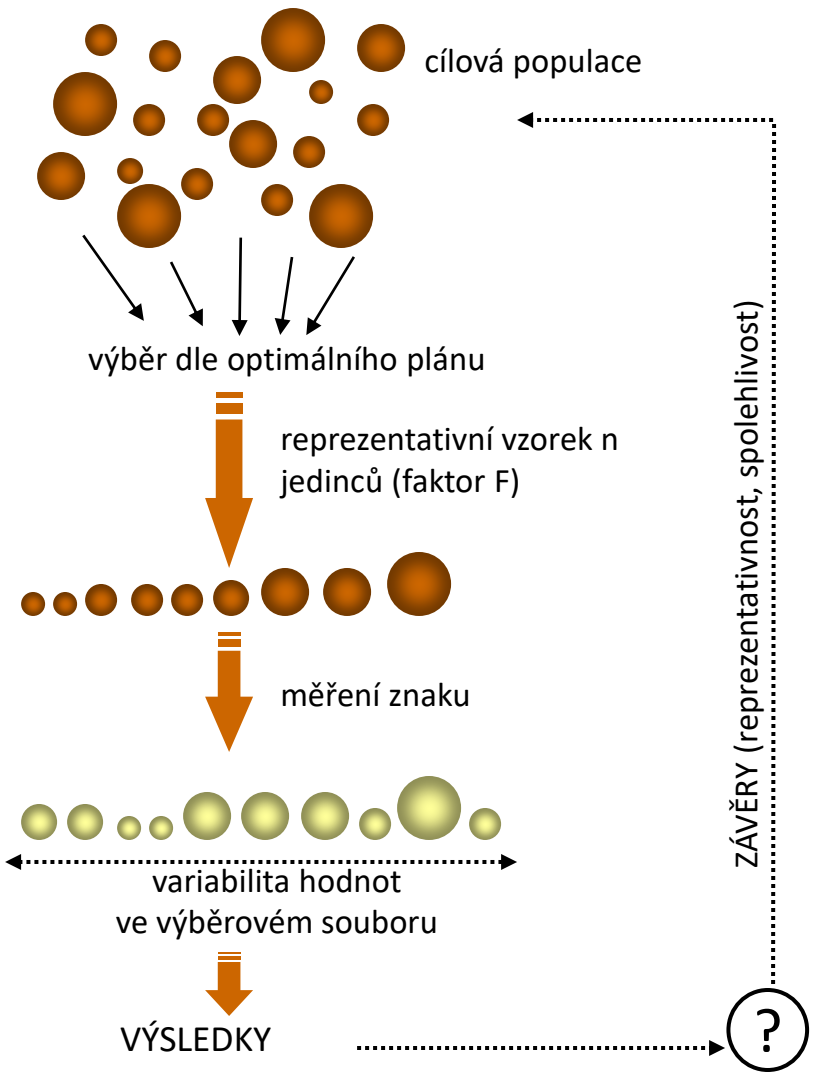
Přístup biostatistiky

- Schopnost: vidět data – komunikovat – interpretovat - prodávat





Experimentální design: nezbytná výbava biologa



Účel analýzy: Popisný

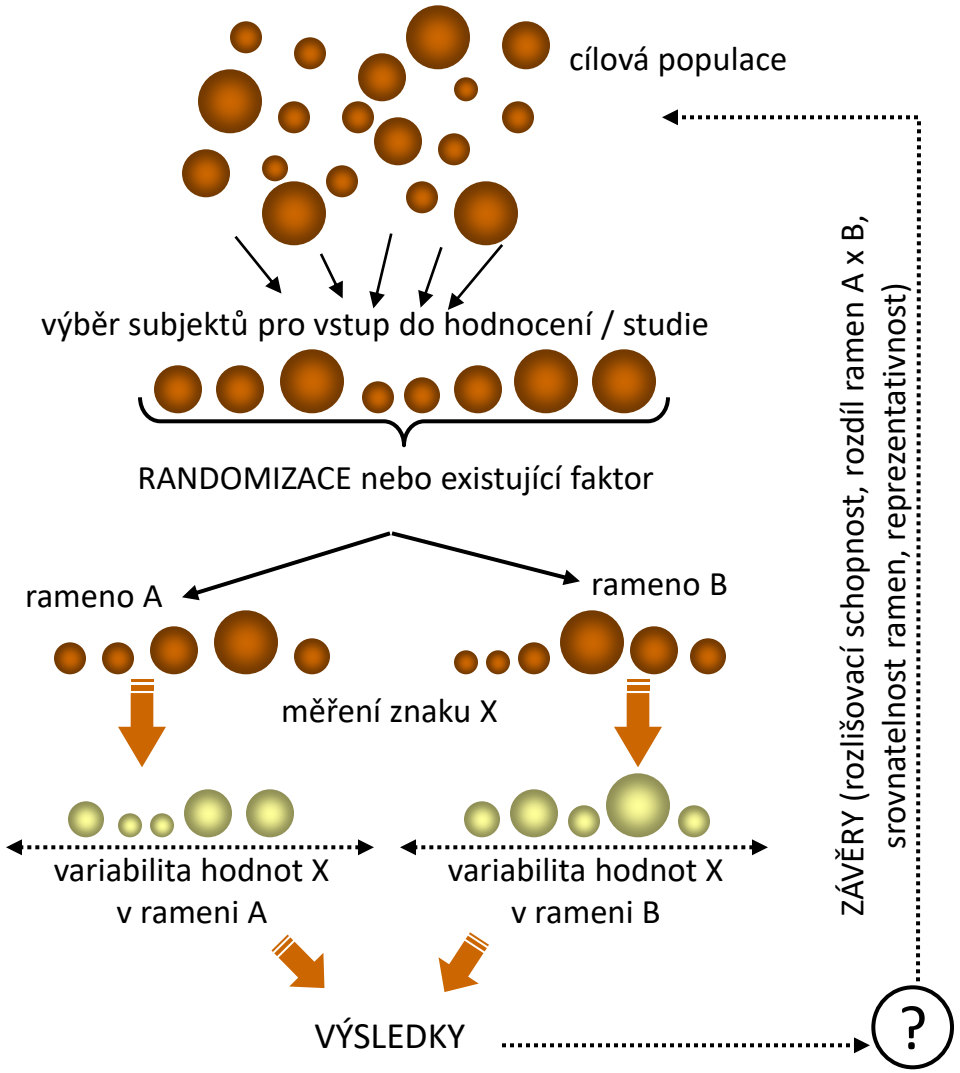
?

Reprezentativnost

Spolehlivost

Přesnost

Experimentální design: nezbytná výbava biologa



Účel analýzy: Srovnávací (2 skupiny)

?

Reprezentativnost

Srovnatelnost

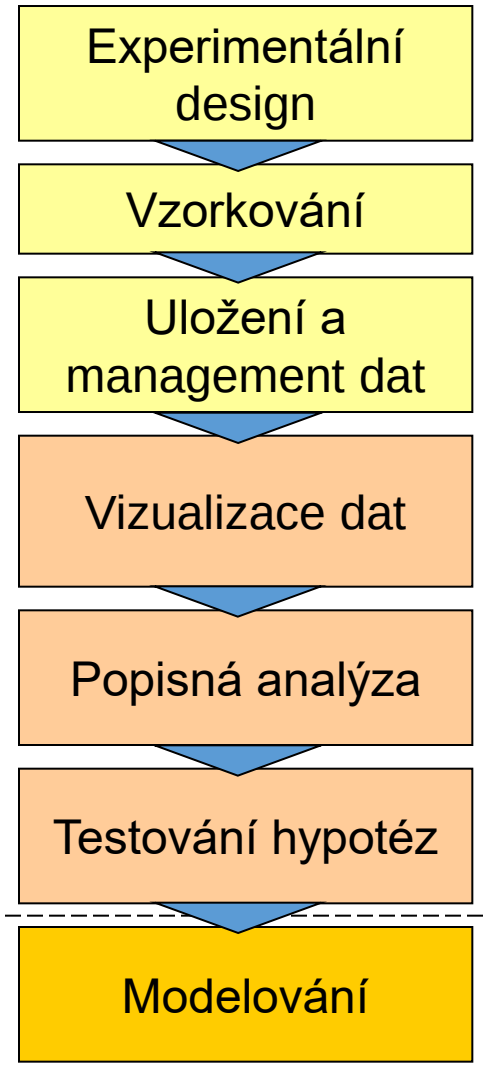
Spolehlivost

Přesnost

… analyzovaný znak cílové populace (X)

… jiný významný faktor charakterizující cílovou populaci (F)

Obecné schéma využití 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.

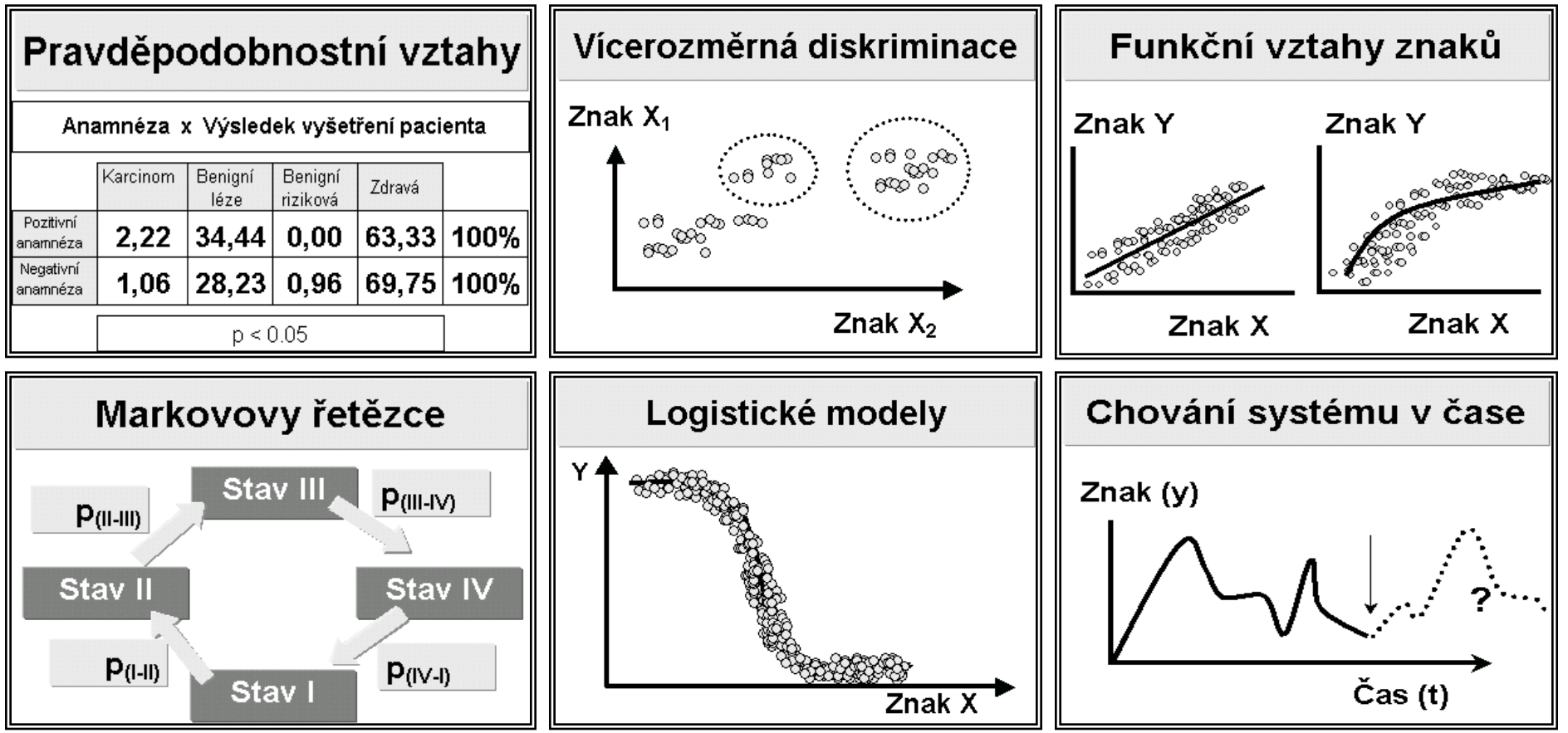
Popisná analýza umožňuje vyhodnotit srovnáním s existující literaturou realističnost naměřených rozsahů dat.

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

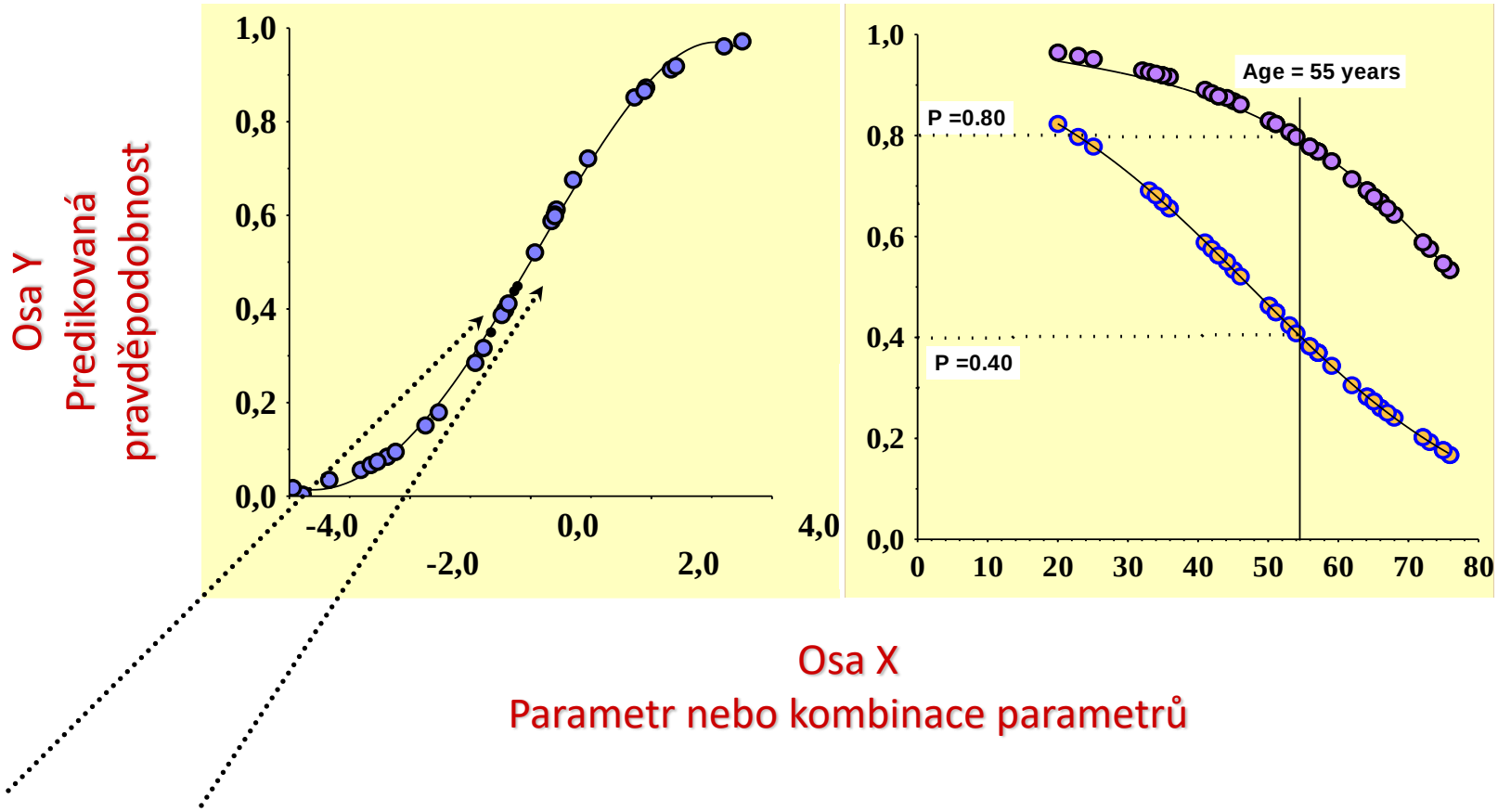
Možným vyvrcholením analýzy je využití získaných znalostí a pochopení problému k vytvoření prediktivních modelů.

Stochastické modelování: predikce neurčitých jevů

- Prospektivně – modelově - postihuje chování jevů při respektování variability



Stochastické modelování: predikce neurčitých jevů



Data konkrétních objektů k přímému hodnocení

Stochastické modelování: predikce neurčitých jevů

- Schopnost: vytvářet prakticky využitelné nástroje