Vícerozměrné statistické metody

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

Jiří Jarkovský, Simona Littnerová

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:

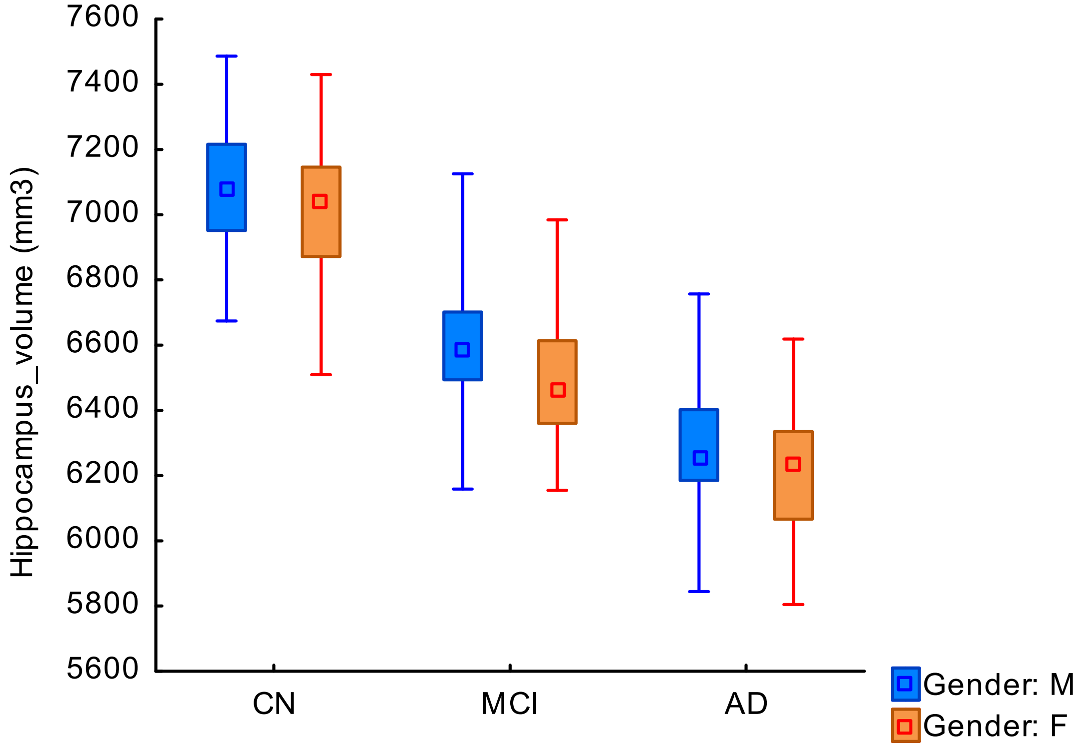
1. Testování hypotéz o vícerozměrných datech
2. Vytvoření shluků subjektů, objektů nebo proměnných
3. Redukce vícerozměrných dat
4. Klasifikace subjektů či objektů
5. Predikce spojitých hodnot

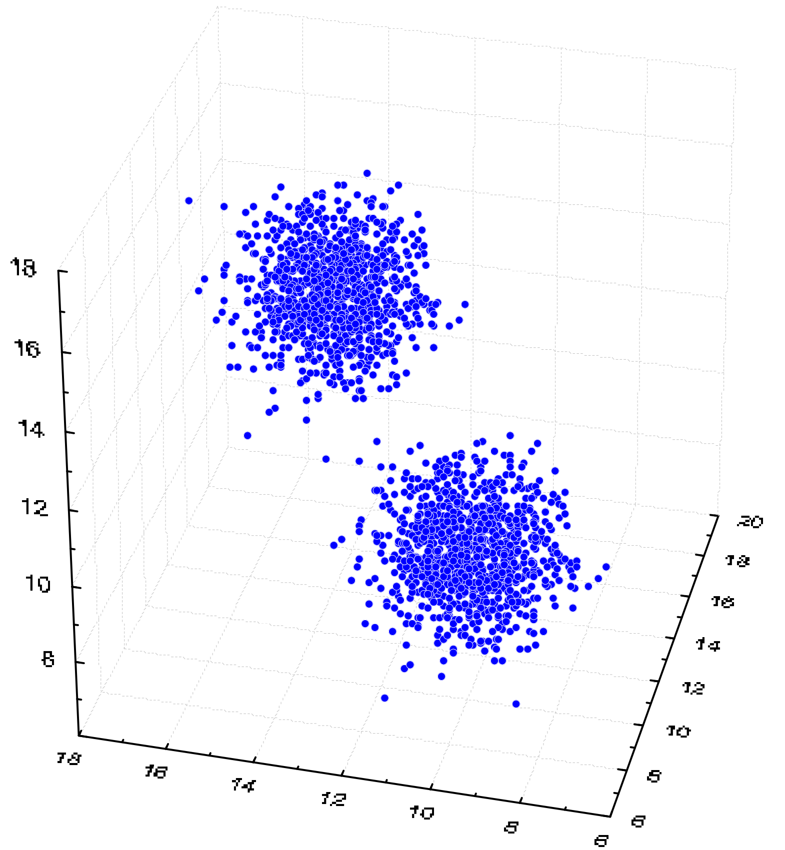
Cíle vícerozměrné analýzy dat

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



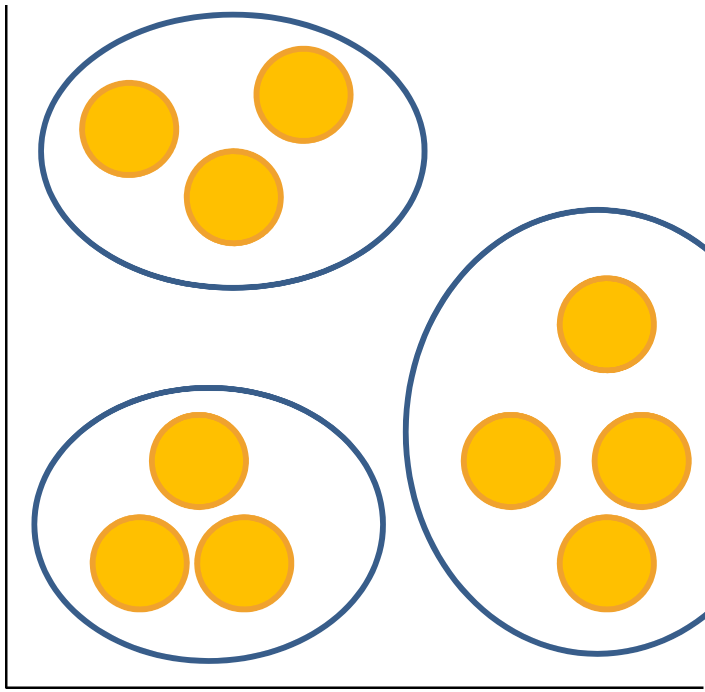


Cíle vícerozměrné analýzy dat

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

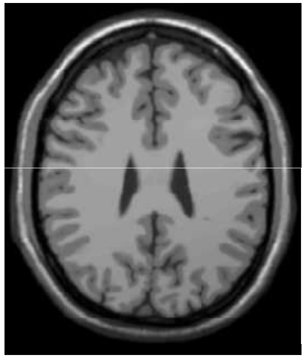
Cíle vícerozměrné analýzy dat

4. Klasifikace subjektů či objektů

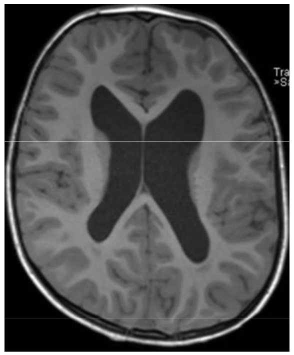
Příklady:

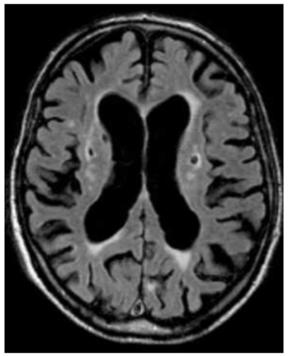
- zjištění (diagnostika) schizofrenie na základě kognitivních testů
- rozhodnutí, zda banka poskytne či neposkytne hypotéku danému subjektu na základě jeho příjmů, rodinné situace atd.
- diagnostika demence (tzn. zařazení nového subjektu do skupiny pacientů či kontrol) podle obrázku mozku

Zdravé subjekty



Pacienti



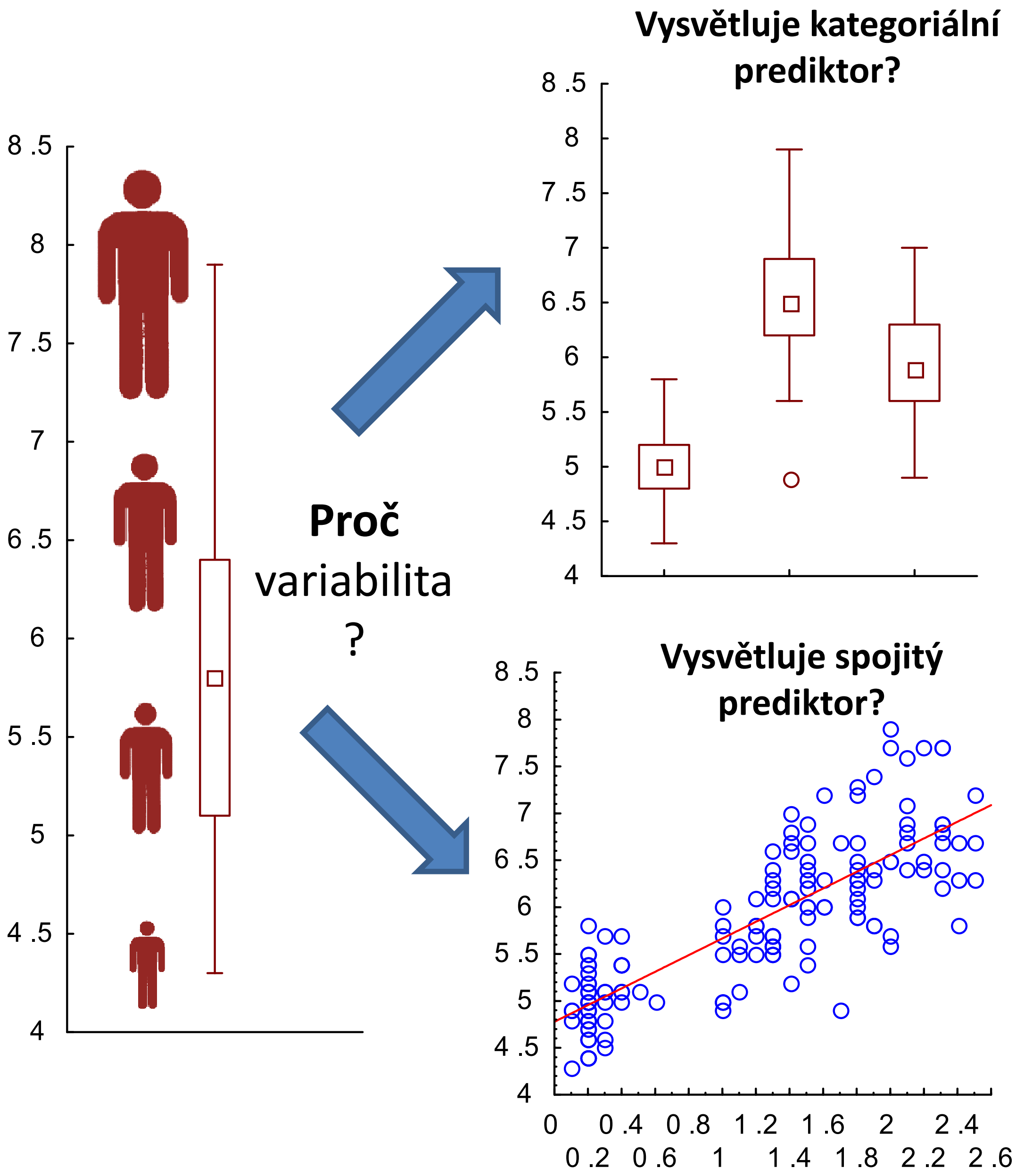


Nový subjekt

Pacient? x Zdravý?

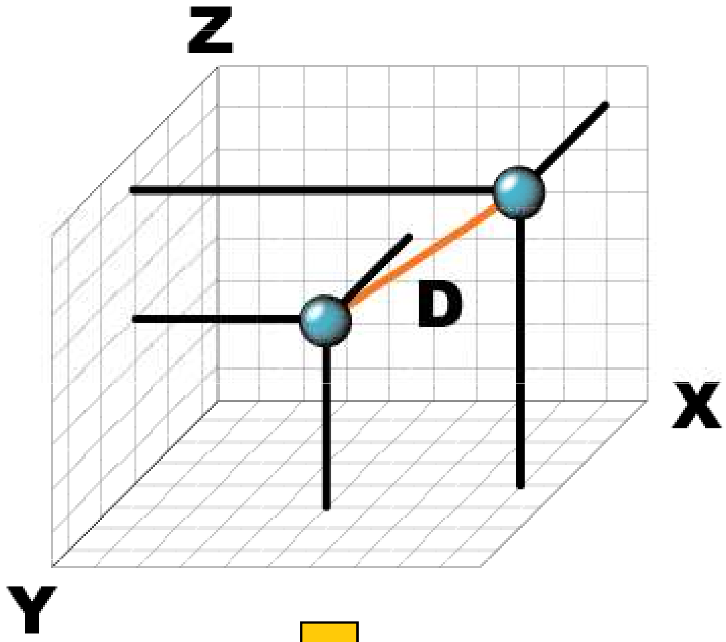
Cíle vícerozměrné analýzy dat
5. Predikce spojitých hodnot

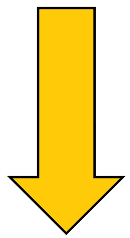
- Obecným cílem je snaha **vysvětlit variabilitu predikované proměnné** (endpoint, Y) pomocí **prediktorů** (vysvětlující proměnná, faktor, X)
- Jak predikovaná proměnná, tak prediktor mohou být různého typu
  - Binární
  - Kategoriální
  - Ordinální
  - Spojitá
  - Cenzorovaná (-> analýza přežití)
- Kombinace datového typu predikované proměnné a prediktoru určuje použitou metodu analýzy



Cíle vícerozměrné analýzy dat - doplnění

- Každý objekt reálného světa můžeme popsat jeho pozicí v mnohorozměrném prostoru, v extrémním případě jde až o desetitisíce 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

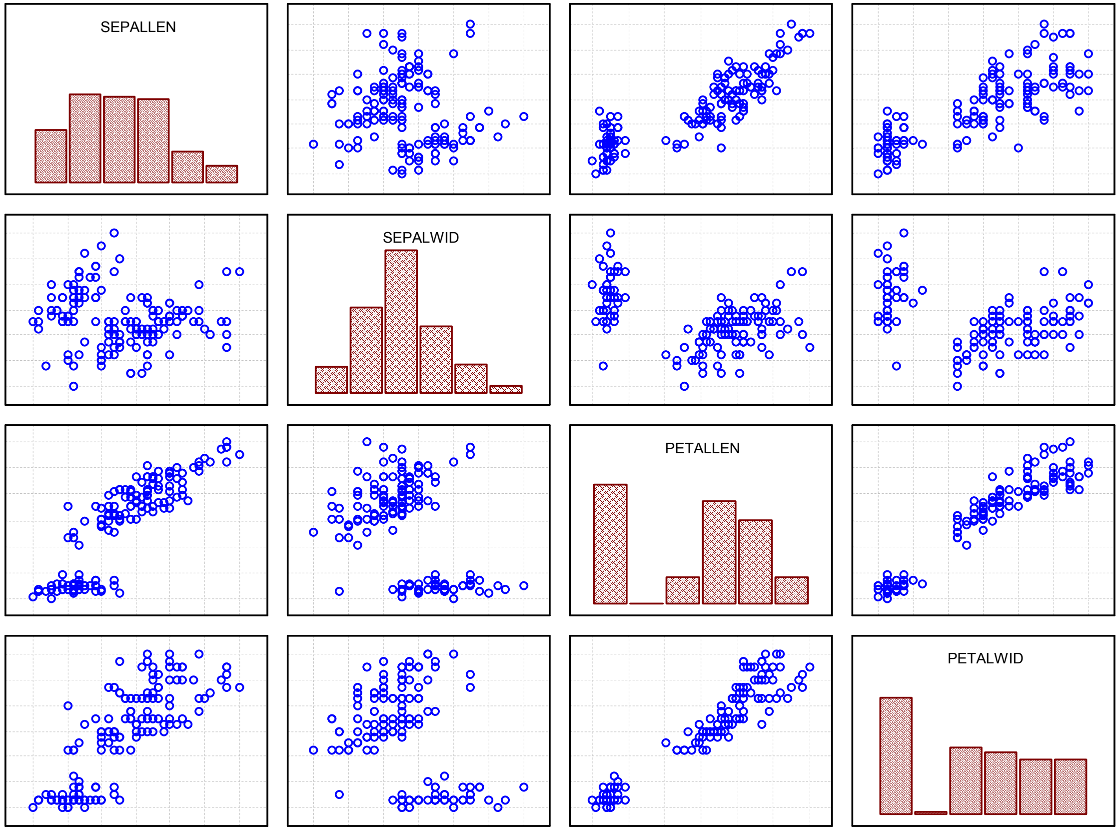




Zjednodušení
Interpretace

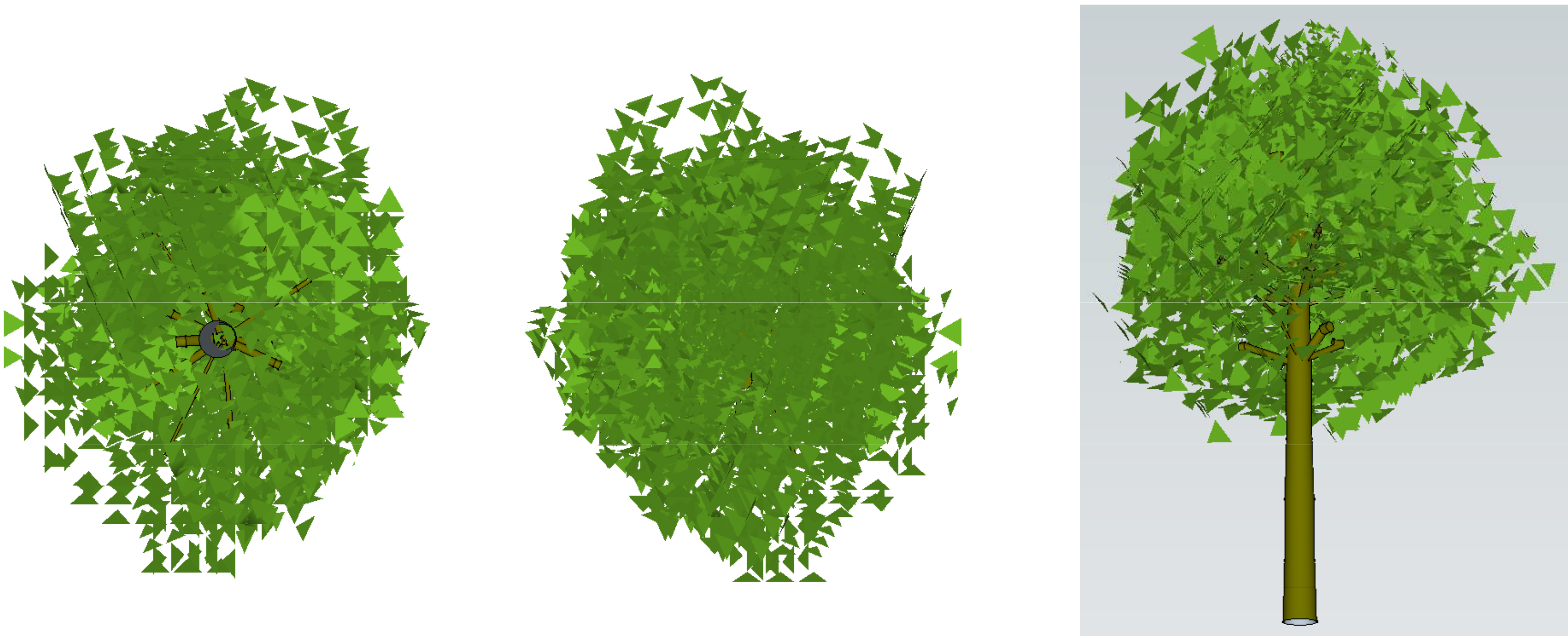
Příklad vícerozměrného popisu objektů

	Dimenze 1	Dimenze 2	Dimenze 3	Dimenze 4
ID objektu	**SEPALLEN**	**SEPALWID**	**PETALLEN**	**PETALWID**
SETOSA	5.0	3.3	1.4	0.2
VIRGINIC	6.4	2.8	5.6	2.2
VERSICOL	6.5	2.8	4.6	1.5
VIRGINIC	6.7	3.1	5.6	2.4
VIRGINIC	6.3	2.8	5.1	1.5
SETOSA	4.6	3.4	1.4	0.3
VIRGINIC	6.9	3.1	5.1	2.3
VERSICOL	6.2	2.2	4.5	1.5
VERSICOL	5.9	3.2	4.8	1.8
SETOSA	4.6	3.6	1.0	0.2
	...	...	...	...



Vícerozměrná analýza dat = pohled ze správného úhlu

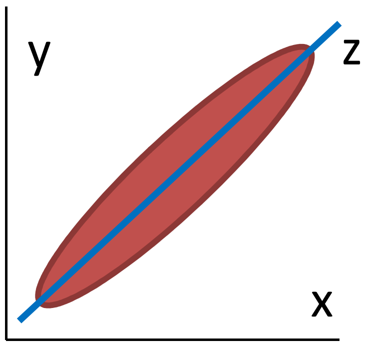
- Vícerozměrná analýza nám pomáhá nalézt v x-dimenzionálním prostoru nejvhodnější pohled na data poskytující maximum informací o analyzovaných objektech



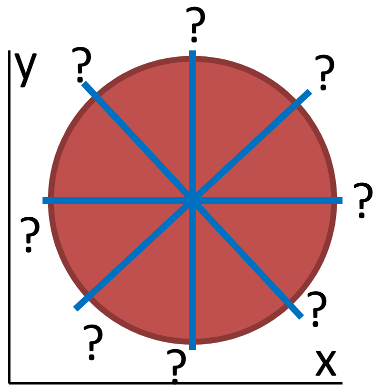
Všechny obrázky ukazují stejný objekt z různých úhlů v 3D prostoru.

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, <u>nemá smysl hledat zjednodušení vícerozměrné struktury takovéhoto souboru !!!</u>



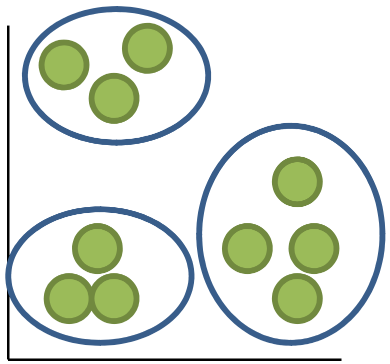
Jednoznačný vztah dimenzí x a y umožňuje jejich nahrazení jedinou novou dimenzí z



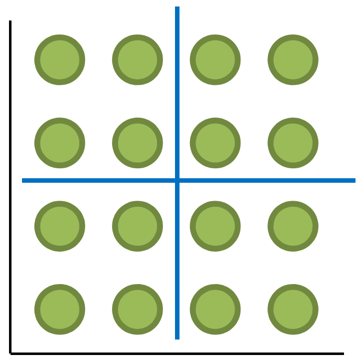
V případě neexistence vztahu mezi x a y nemá smysl definovat nové dimenze – nepřináší žádnou novou informaci oproti x a y

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)



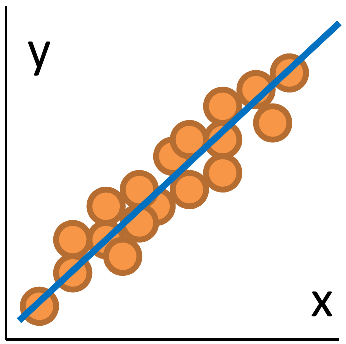
Shluková analýza je možná i v tomto případě, nicméně hranice shluků jsou dány pouze naším rozhodnutím.

Omezení vícerozměrné analýzy dat

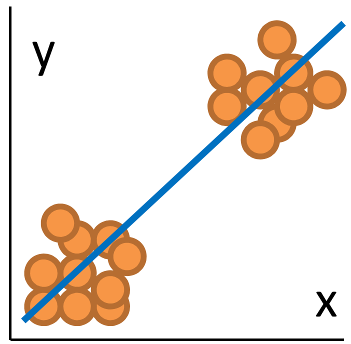
- Vícerozměrná analýza může přinést zjednodušení dimenzionality dat pouze v případě, kdy data skrývají nějakou identifikovatelnou vícerozměrnou strukturu
  - Mezi dimenzemi existují vztahy (korelace) umožňující nahrazení korelovaných dimenzí zástupnou souhrnnou dimenzí
  - Objekty vytváří v x-dimenzionálním 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
- 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:
  - Vybrána vhodná metoda 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

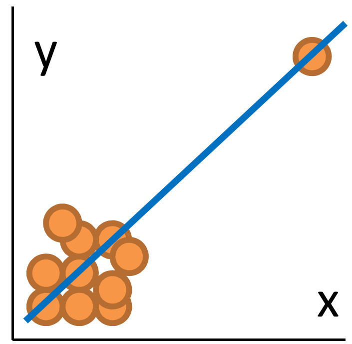
- Kovariance a Pearsonova korelace je základem analýzy hlavních komponent, faktorové analýzy jakož i dalších vícerozměrných analýz pracujících s lineární závislostí proměnných
- Předpokladem výpočtu kovariance a Pearsonovy korelace je:
  - Normalita dat v obou dimenzích
  - Linearita vztahu proměnných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



Lineární vztah – bezproblémové použití Personovy 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 tabule 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 chi-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

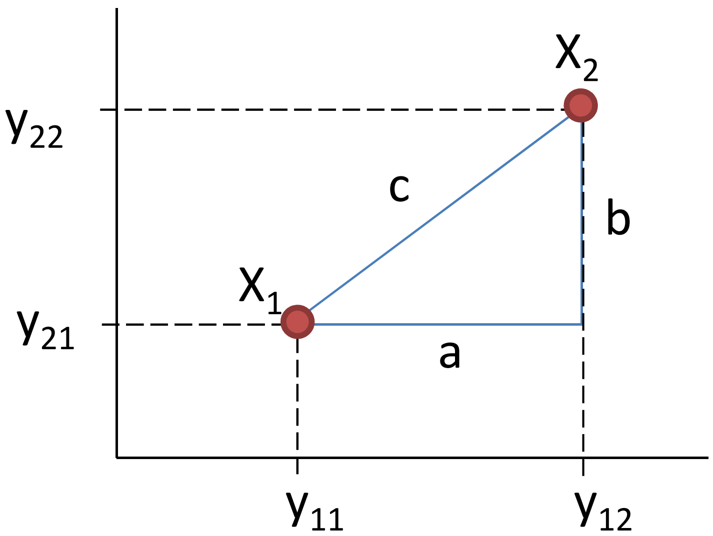
	☠	☺
A	5	5
B	5	5

Očekávaná tabulka

Hodnota chi-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

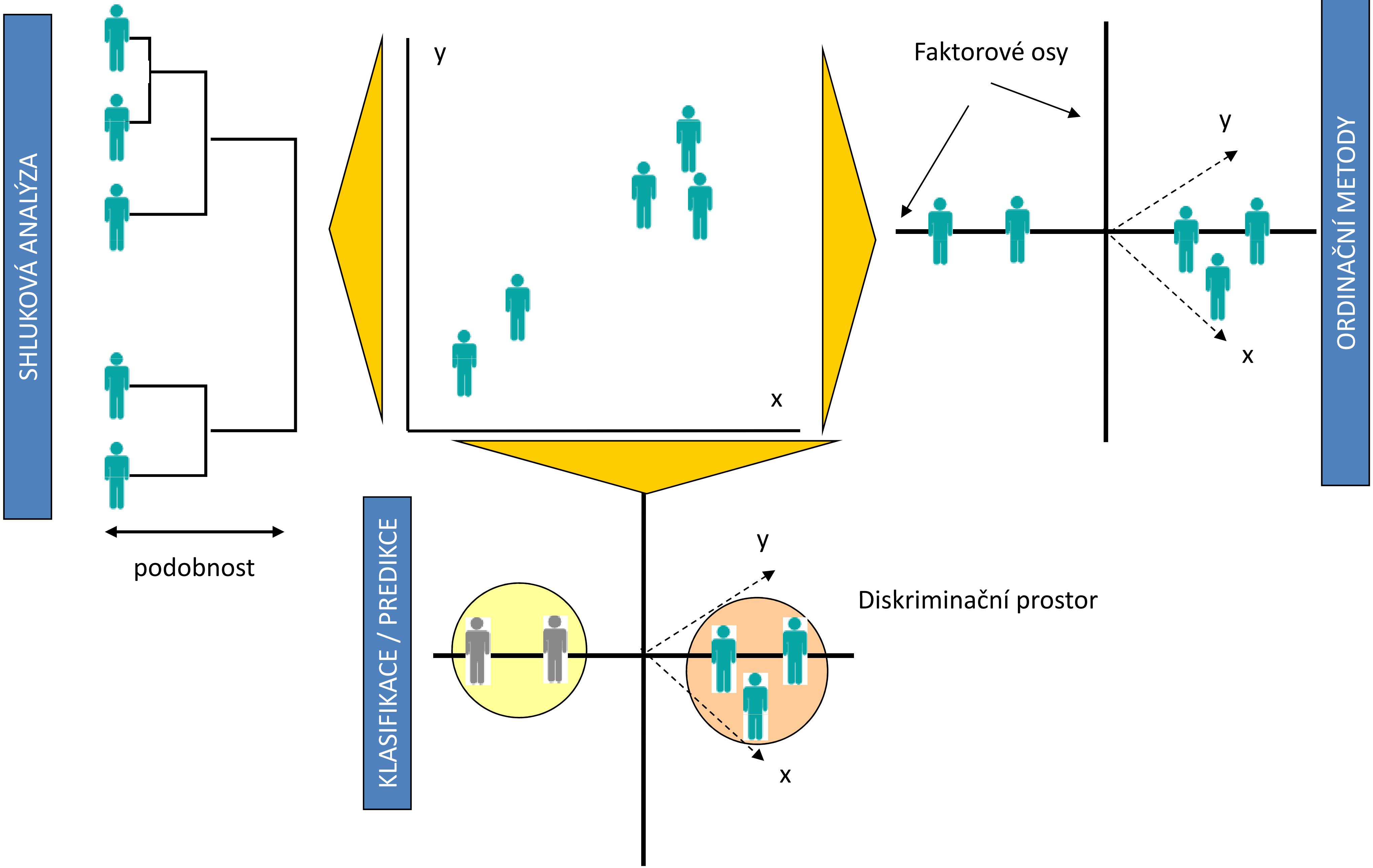
KLASIFIKACE / PREDIKCE

- Na základě vícerozměrné kombinace prediktorů zařazujeme objekty do skupin (klasifikace) nebo predikujeme spojitou proměnnou (predikce)

ORDINAČNÍ METODY

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

Typy vícerozměrných analýz

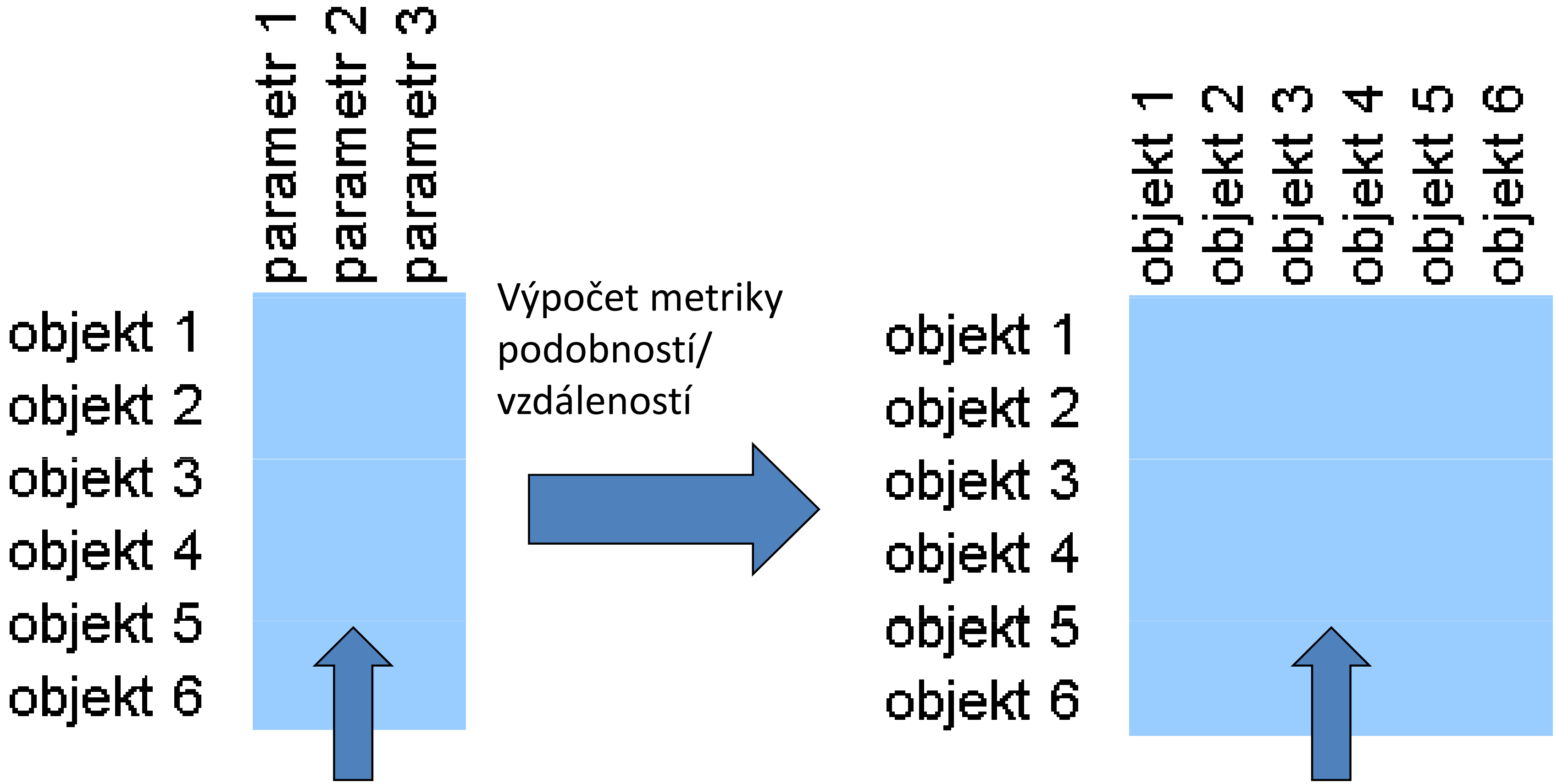


Pojmy vícerozměrných analýz

- 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.

Vstupní matice vícerozměrných analýz

NxP MATICE **ASOCIAČNÍ MATICE**



Hodnoty parametrů pro jednotlivé objekty

Korelace, kovariance, vzdálenost, podobnost

Vícerozměrné statistické metody

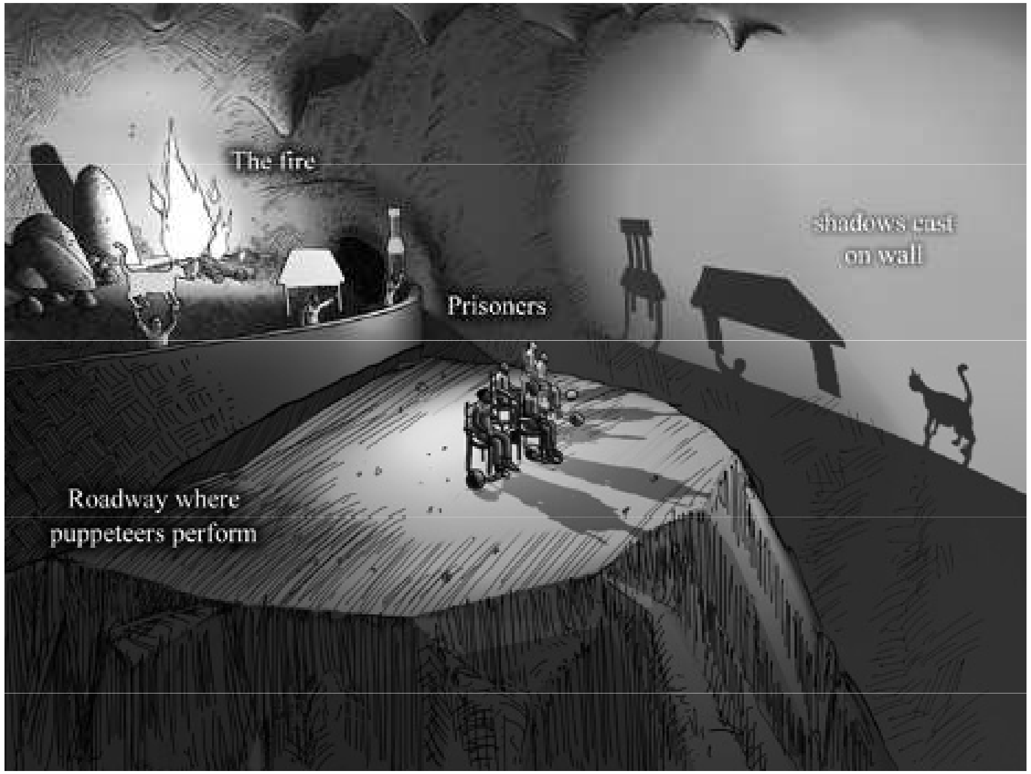
Jednorozměrná statistická analýza jako předpoklad vícerozměrné analýzy dat

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

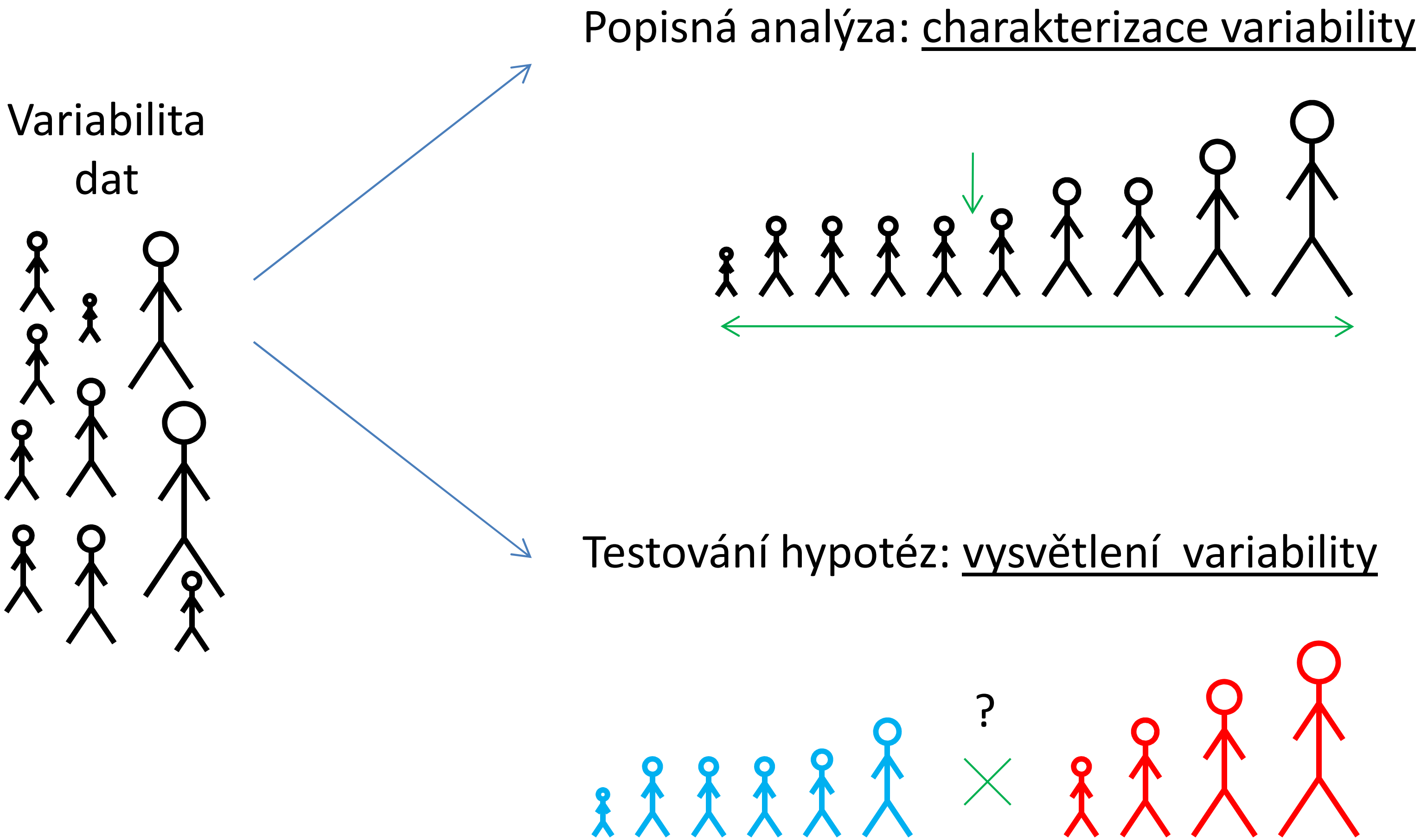


Variabilita jako základní pojem ve statistice

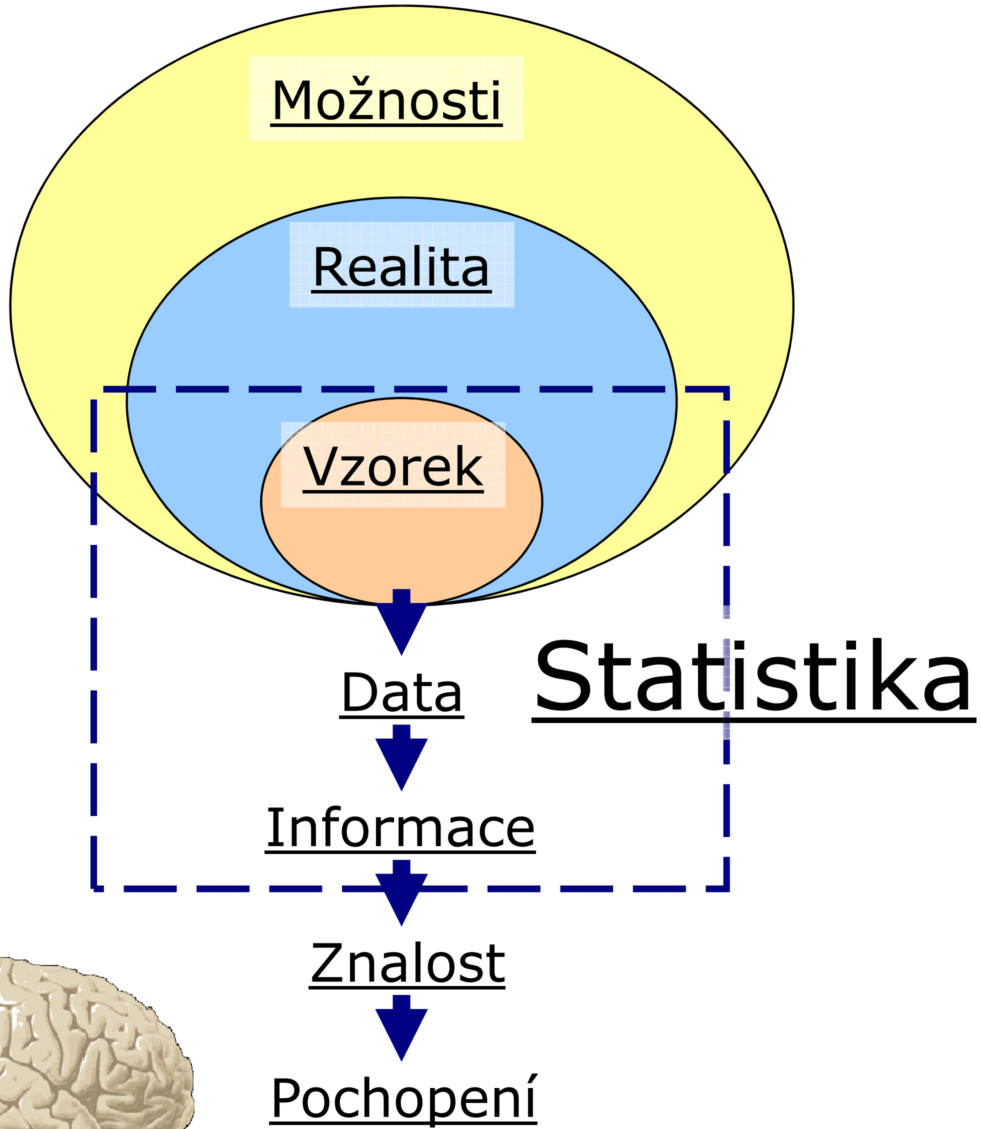
- Naše realita je variabilní a statistika je vědou zabývající se variabilitou
- Korektní analýza variabilita a její pochopení přináší užitečné informace o naší realitě
- V případě deterministického světa by statistická analýza nebyla potřebná

Práce s variabilitou v analýze dat

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

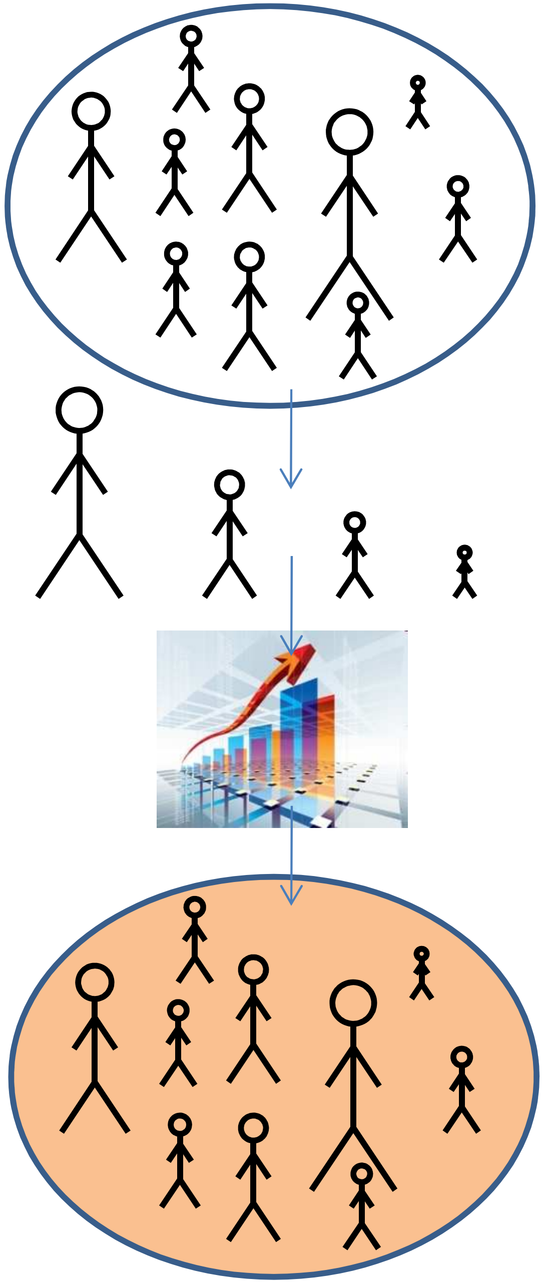


Co může statistika říci o naší realitě?



- 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 !!!

Statistika a zobecnění výsledků



Neznámá cílová populace

Vzorek

Analýza

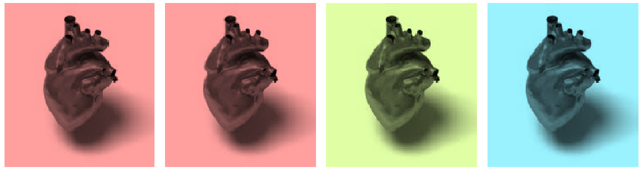
Díky zobecnění výsledků známe vlastnosti cílové populace

- 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 je nutné dodržet

- Náhodný výběr z cílové populace
- Representativnost: struktura vzorku musí maximálně reflektovat realitu



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

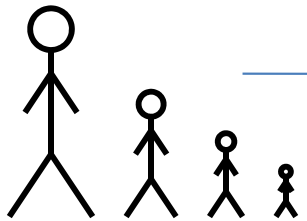
Velikost vzorku a přesnost statistických výstupů

- Existuje skutečné rozložení a skutečný průměr měřené proměnné

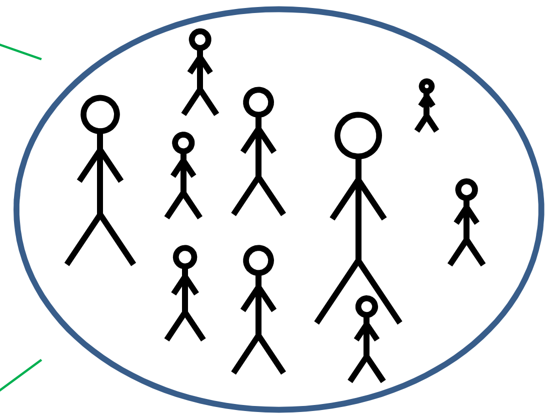
- Z jednoho měření nezjistíme nic

???

- Vzorek určité velikosti poskytuje odhad reálné hodnoty s definovanou spolehlivostí



Odhad průměru atd.



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

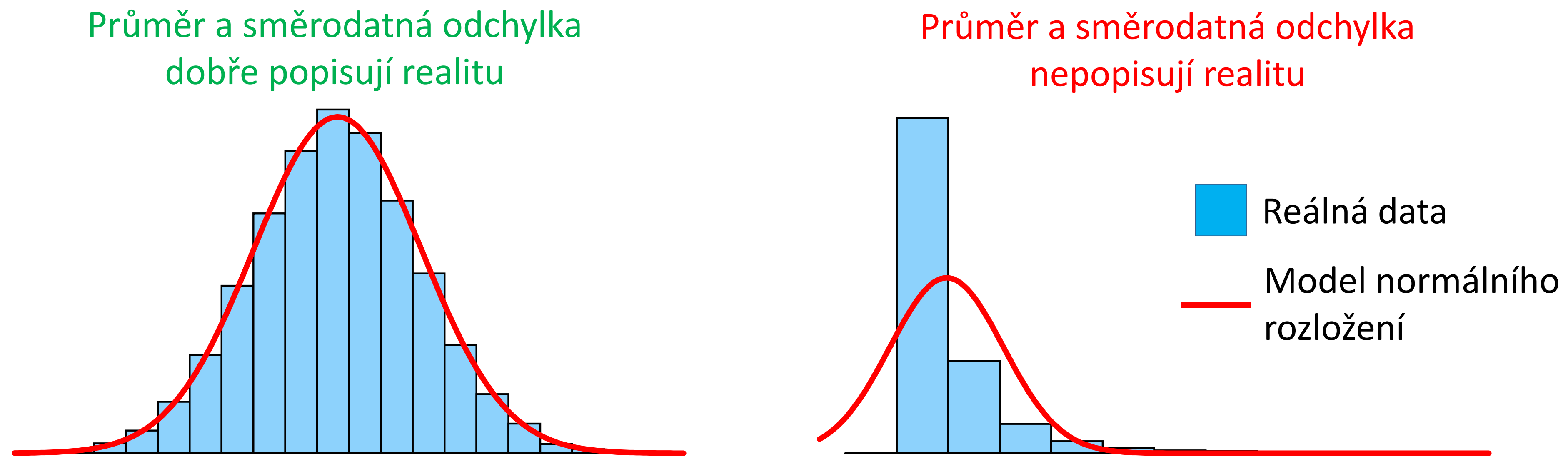
Předpoklady statistické analýzy

- 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.
- <u>Statistika funguje korektně pouze pokud jsou splněny předpoklady jejích metod a modelů.</u>

Normální rozložení jako předpoklad statistické analýzy dat

- Normální rozložení (Gaussova křivka) je jedním z hlavních modelů ve statistické analýze dat
- Řada metod popisné statistiky je založena na modelu normálního rozložení
  - Průměr, směrodatná odchylka atd.
- Řada metod testování hypotéz je založena na modelu normálního rozložení
  - T-test, ANOVA, korelace, regrese



- Použití modelu je možné pouze pokud reálná data odpovídají danému modelovému rozložení

Obecné schéma aplikace statistické analýzy

Experimentální design	Jak velký vzorek je nezbytný pro statisticky relevantní výsledky? Klíčová stratifikační kritéria cílové populace.
Vzorkování	Vzorkovací plán zabezpečující náhodnost a reprezentativnost vzorku.
Uložení a management dat	Uložení dat ve vhodné formě a jejich vyčištění předcházející vlastní analýze je klíčovým krokem statistické analýzy.
Vizualizace dat	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	Popisná analýza umožňuje vyhodnotit srovnáním s existující literaturou realističnost naměřených rozsahů dat.
Testování hypotéz	Testování vazeb mezi různými proměnnými s cílem navzájem vysvětlit jejich variabilitu a tím přispět k pochopení řešeného problému.
Modelování	Možným vyvrcholením analýzy je využití získaných znalostí a pochopení problému k vytvoření prediktivních modelů.

Vícerozměrné statistické metody

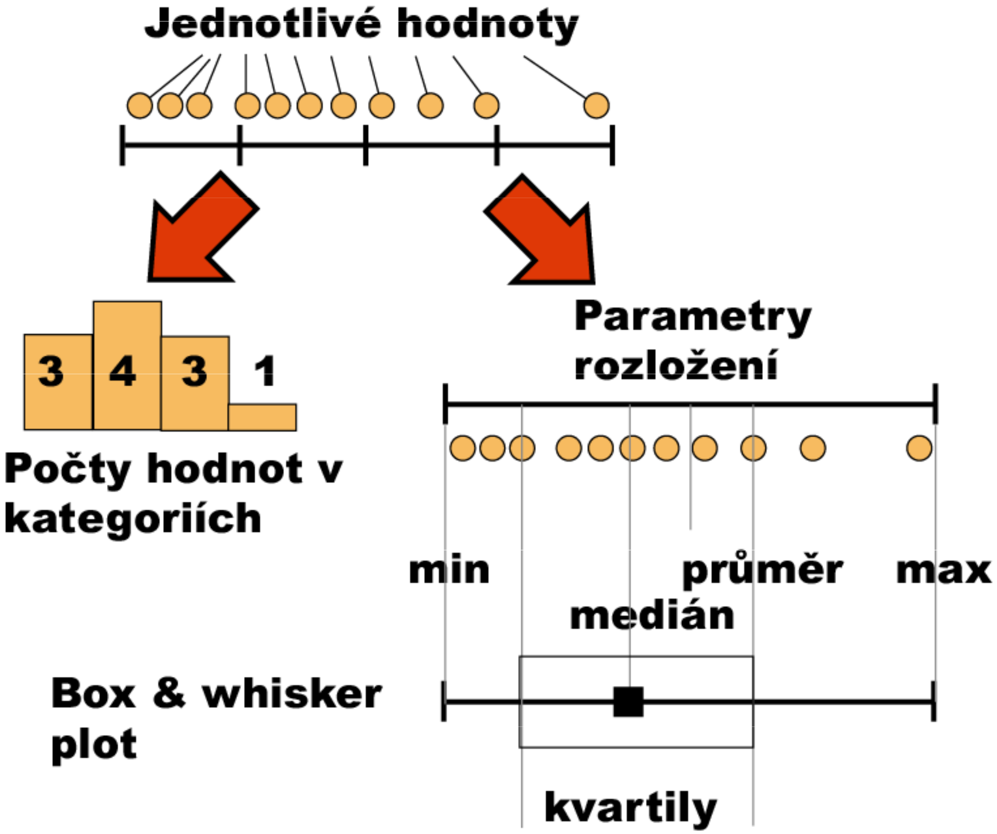
Popisná statistika a její spolehlivost

Typy proměnných a jejich popisné statistiky

- Kvalitativní/kategorická
  - binární - ano/ne
  - nominální - A,B,C … několik kategorií
  - ordinální - 1<2<3 …několik kategorií a můžeme se ptát, která je větší
  - **Popis procentuálním zastoupením kategorií**

- Kvantitativní
  - nespojitá – čísla, která však nemohou nabývat všech hodnot (např. počet porodů)
  - spojitá – teoreticky jsou možné všechny hodnoty (např. krevní tlak)
  - **Popis celou řadou deskriptivních statistik (průměr, medián, percentily, směrodatná odchylka, rozsah hodnot apod.)**

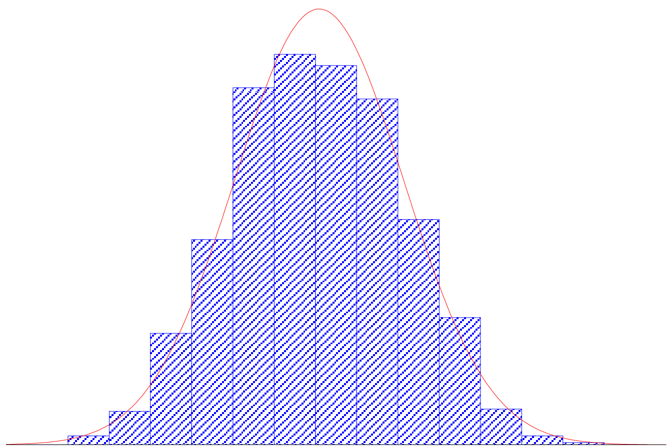
Řada dat a její vlastnosti



Kvalitativní data
Tabulka s četností jednotlivých kategorií.

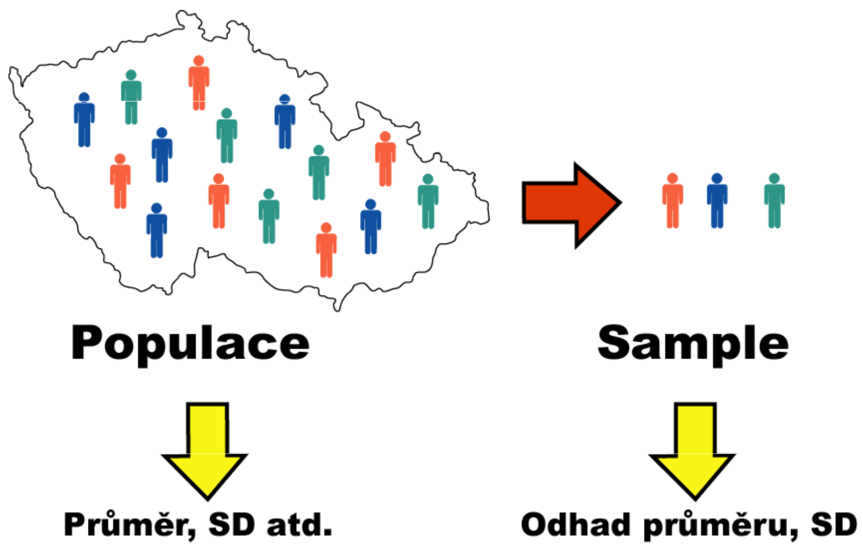
Kategorie	Četnost
B	5
C	8
D	1

Kvantitativní data
Četnost hodnot rozložení v jednotlivých intervalech.



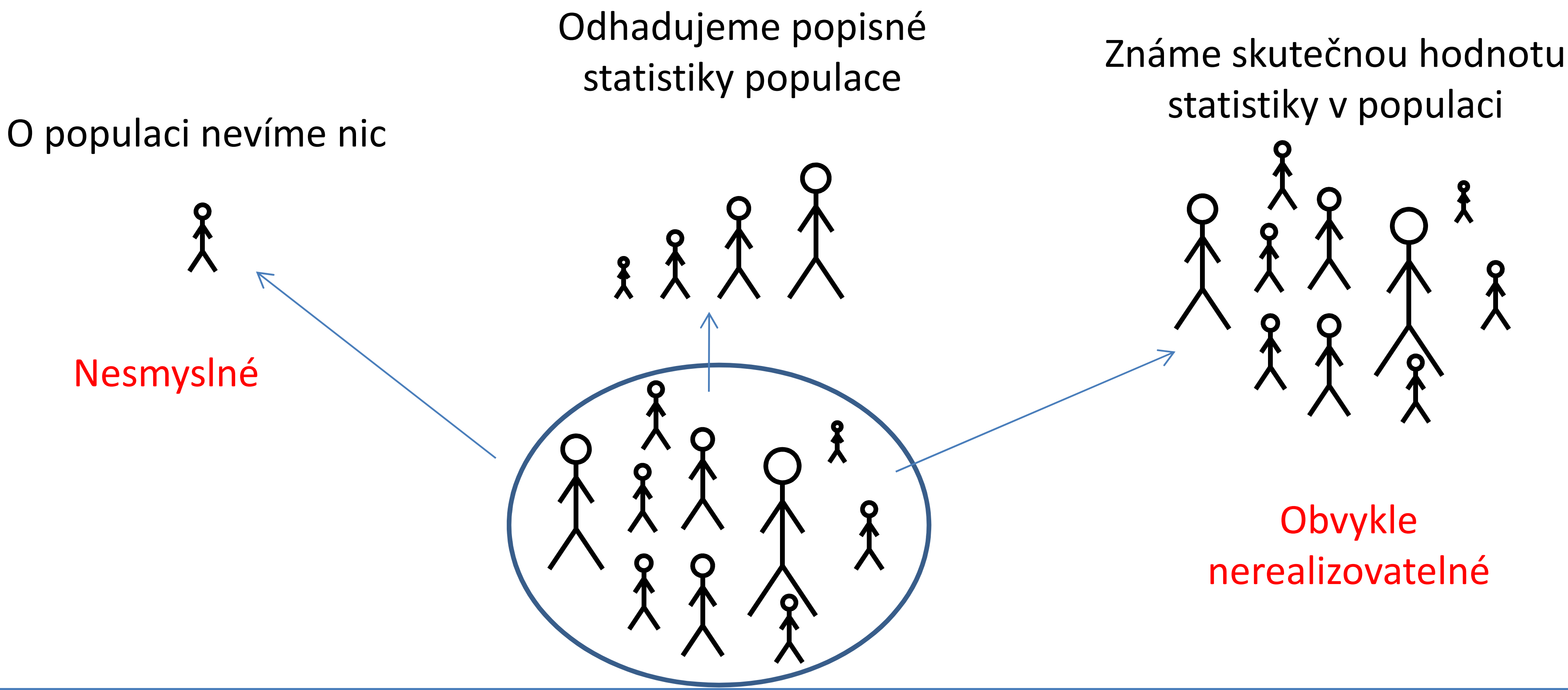
Populace a vzorek

- Populace představuje veškeré možné objekty vzorkování, např. veškeré obyvatelstvo ČR při sledování na úrovni ČR, z populace získáme reálné parametry rozložení
- Z populace je prováděno vzorkování za účelem získání reprezentativního vzorku (sample) populace, toto vzorkování by mělo být náhodné, důležitá je také velikost vzorku, ze vzorku získáme odhady parametrů rozložení



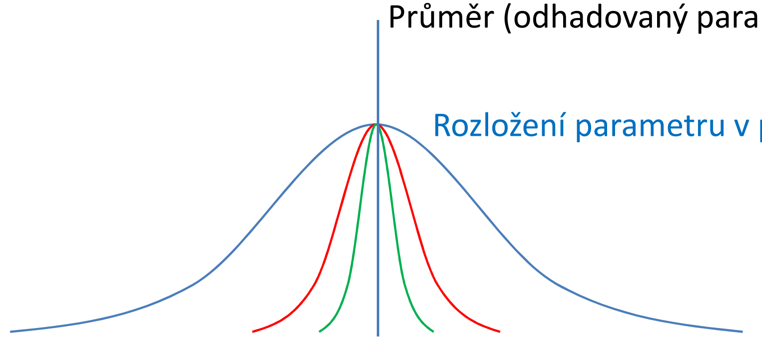
Popisná statistika: odhad reality

- 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



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



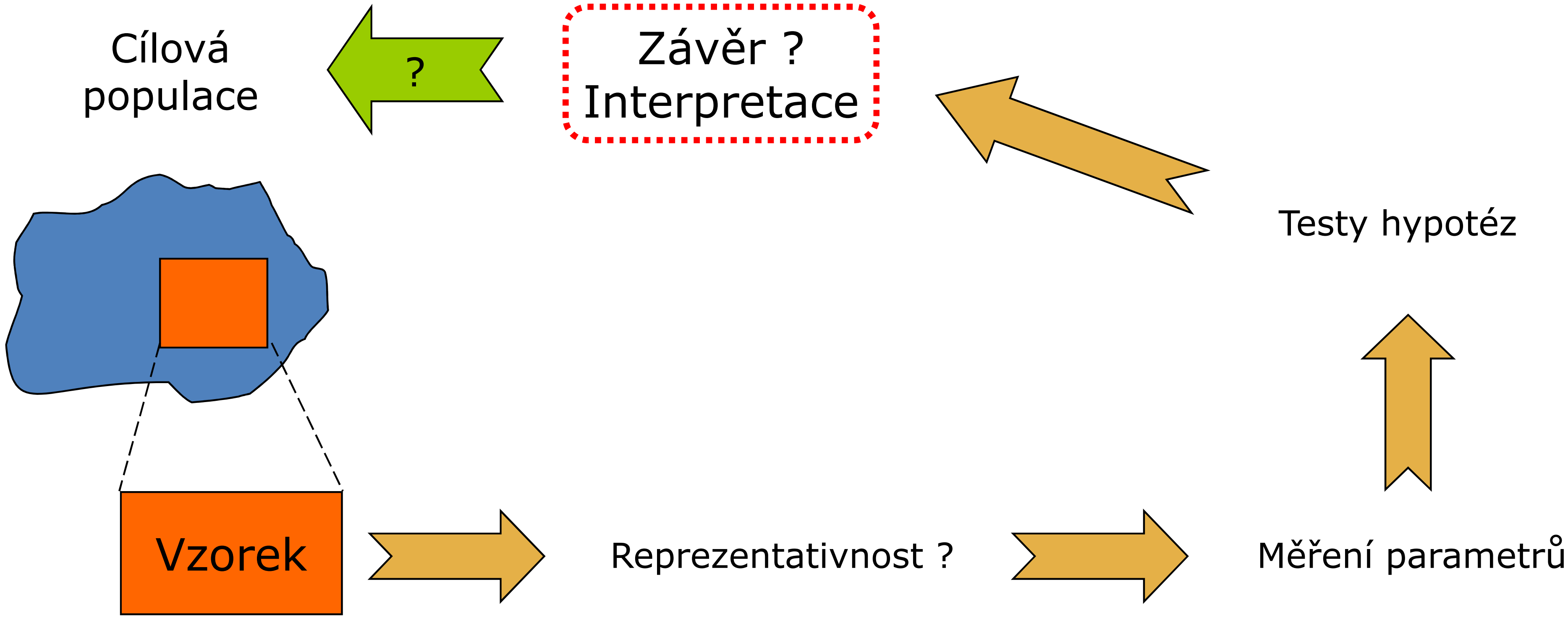
- Interval spolehlivosti lze spočítat pro jakoukoliv statistiku (průměr, směrodatná odchylka, korelace, procentuální zastoupení apod.)
- Interval spolehlivosti poskytuje vodítko jak „spolehlivé“ jsou naše výsledky a s jakou pravděpodobností jich je možné opakovaně dosáhnout
- 95% interval spolehlivosti 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 !!!

Vícerozměrné statistické metody

Testování hypotéz

Testování hypotéz: základní principy

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

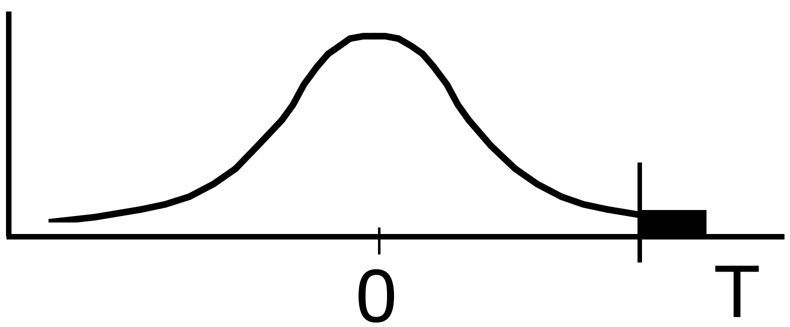


Statistické testování – základní pojmy

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

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

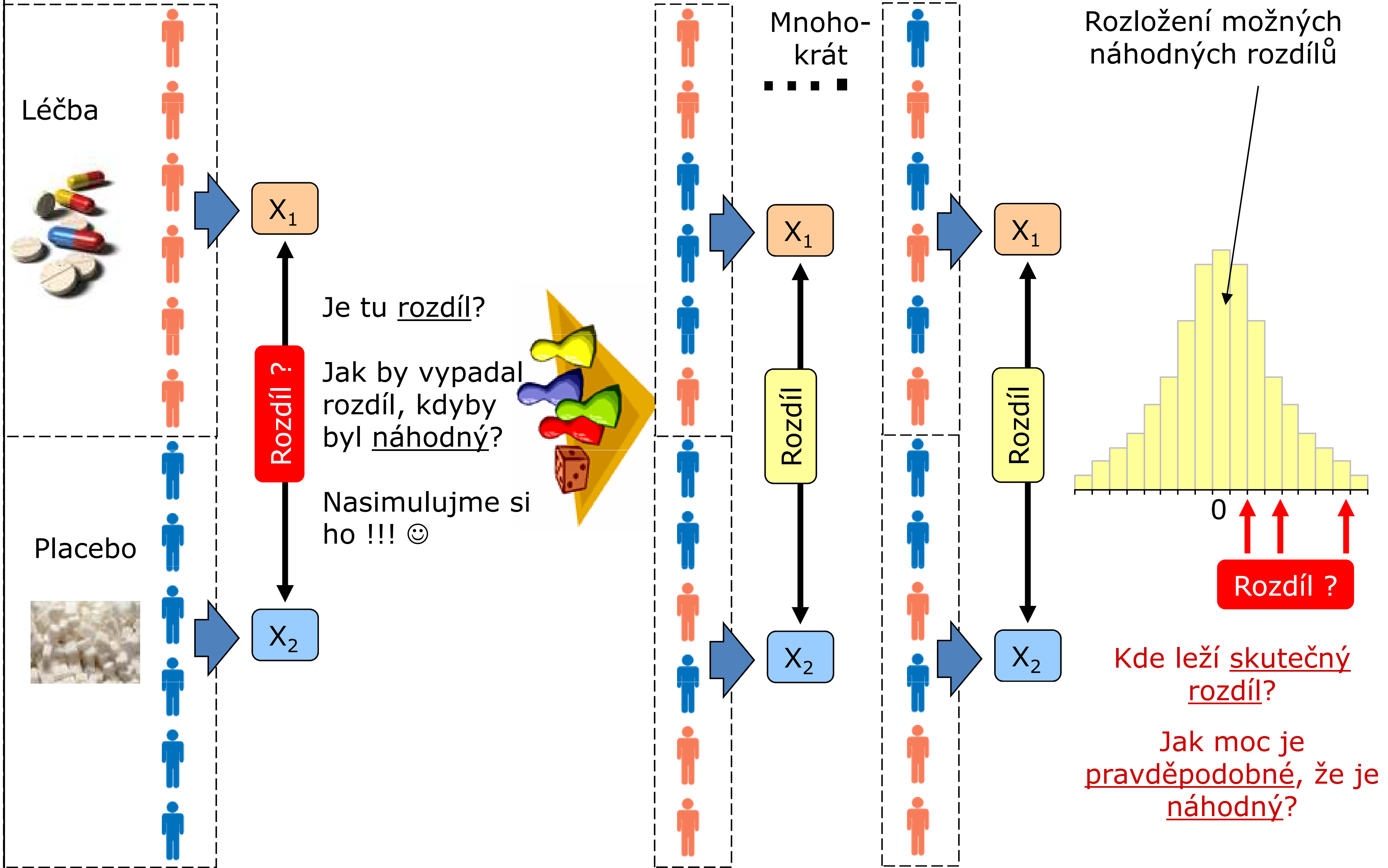
- Kritický obor testové statistiky



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

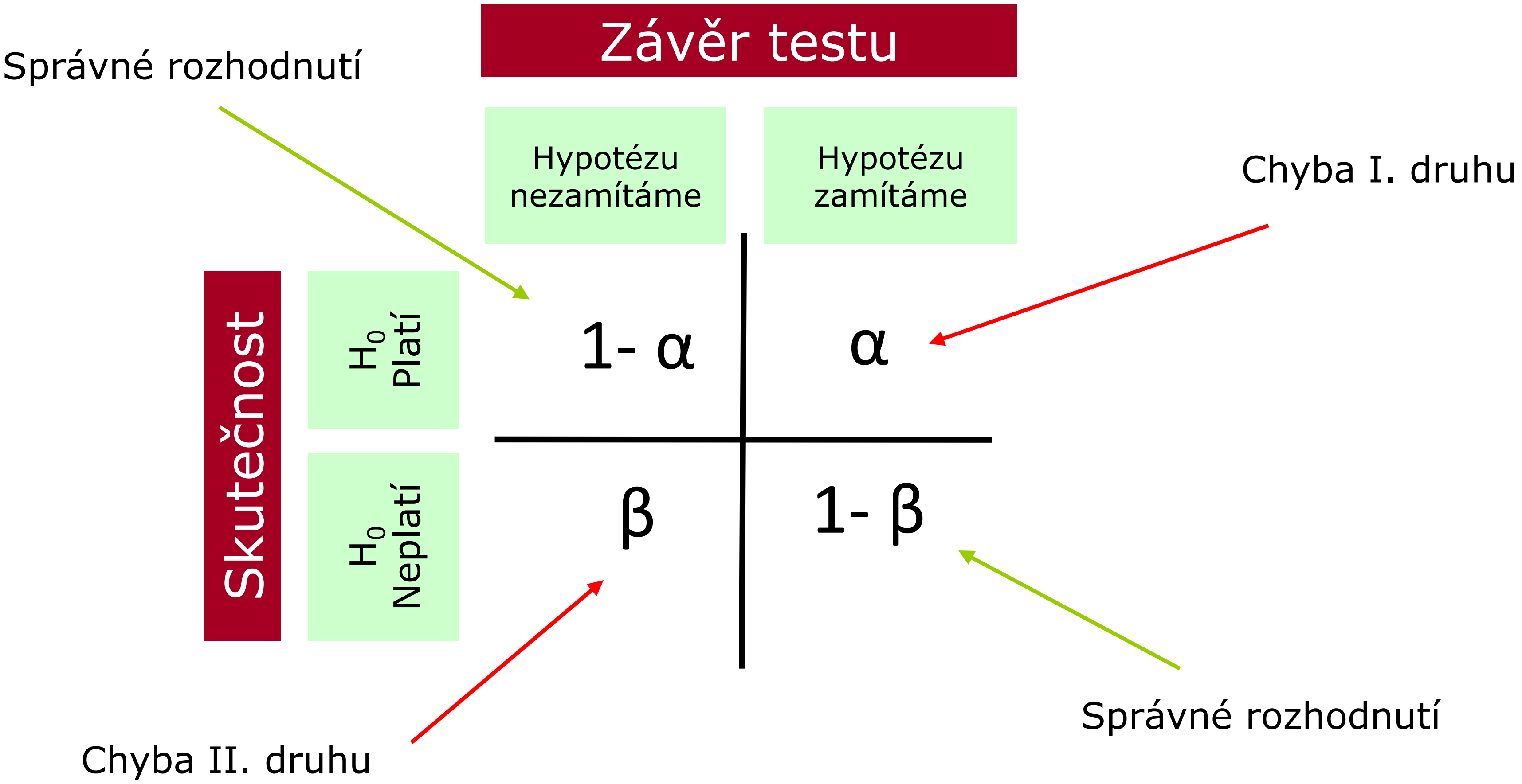
- Statistická významnost (p) – odvozena z testové statistiky a znamená pravděpodobnost, že pozorovaný rozdíl je výsledkem pouhé náhody

Co znamená pravděpodobnost, že pozorovaný rozdíl je výsledkem pouhé náhody ?



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.**

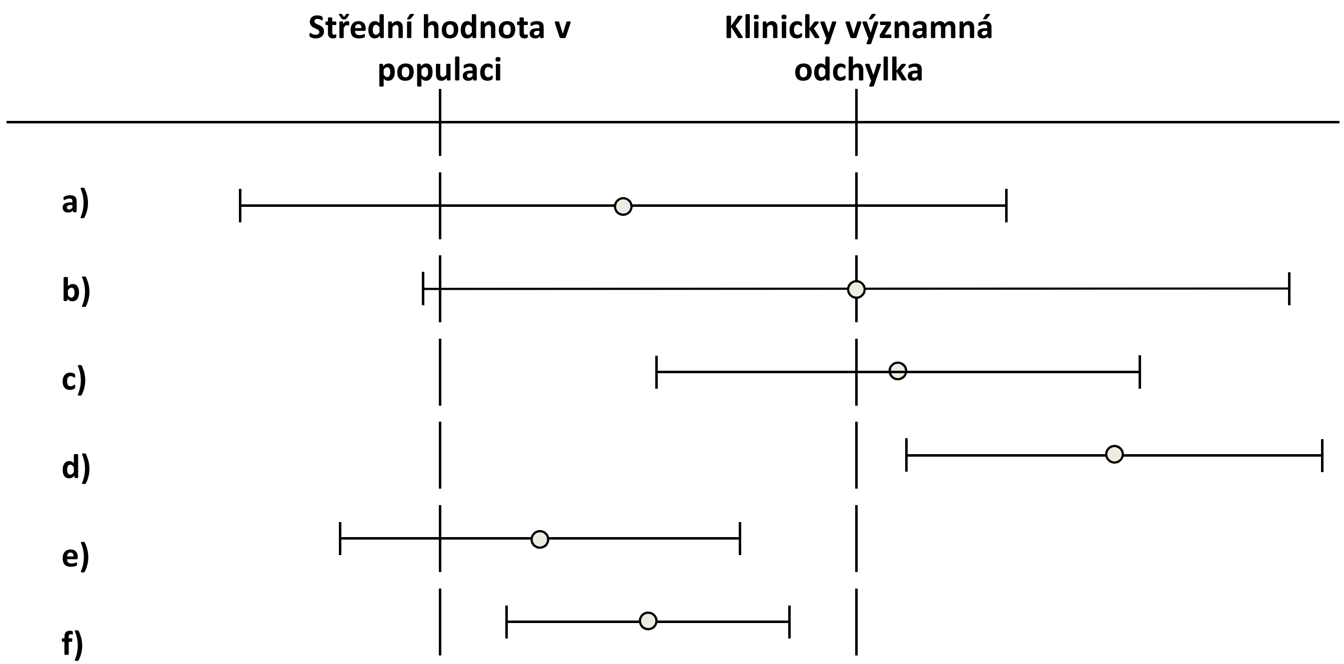


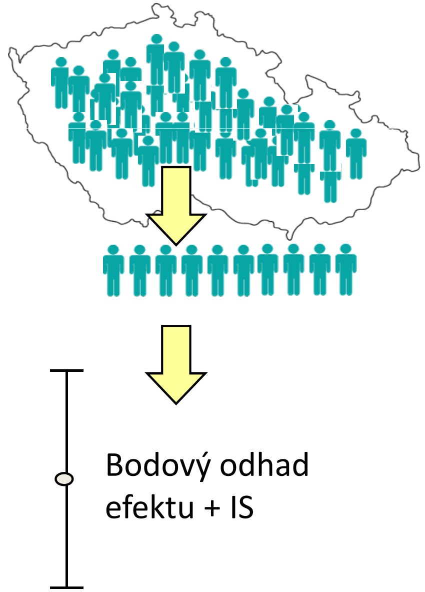
Klinická a statistická významnost

- 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

Statistická vs. klinická významnost





Možnost	Statistická významnost	Klinická významnost
a)	ne	možná
b)	ne	možná
c)	ano	možná
d)	ano	ano
e)	ne	ne
f)	ano	ne

Parametrické vs. neparametrické testy

Parametrické testy

- Mají předpoklady o rozložení vstupujících dat (např. normální rozložení)
- Při stejném N a dodržení předpokladů mají vyšší sílu testu než testy neparametrické
- Pokud nejsou dodrženy předpoklady parametrických testů, potom jejich síla testu prudce klesá a výsledek testu může být zcela chybný a nesmyslný

Neparametrické testy

- Nemají předpoklady o rozložení vstupujících dat, lze je tedy použít i při asymetrickém rozložení, odlehlých hodnotách, či nedetekovatelném rozložení
- Snížená síla těchto testů je způsobena redukcí informační hodnoty původních dat, kdy neparametrické testy nevyužívají původní hodnoty, ale nejčastěji pouze jejich pořadí

One-sample vs. two sample testy

One – sample testy

- Srovnávají jeden vzorek (one sample, jednovýběrové testy) s referenční hodnotou (popřípadě se statistickým parametrem cílové populace)
- V testu je tedy srovnáváno rozložení hodnot (vzorek) s jediným číslem (referenční hodnota, hodnota cílové populace)
- Otázka položená v testu může být vztažena k průměru, rozptylu, podílu hodnot i dalším statistickým parametrům popisujícím vzorek

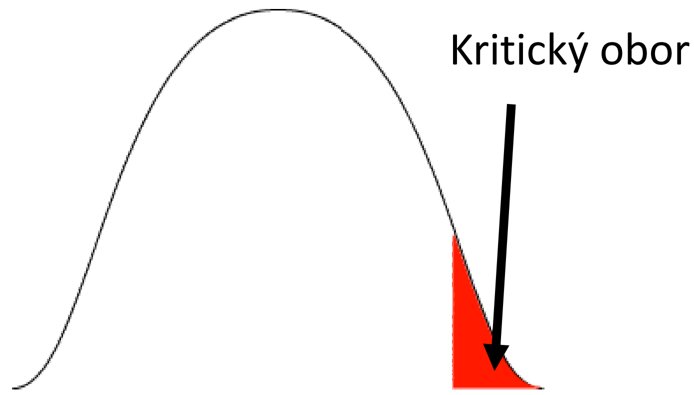
Two – sample testy

- Srovnávají navzájem dva vzorky (two sample, dvouvýběrové vzorky)
- V testu jsou srovnávány dvě rozložení hodnot
- Otázka položená v testu může být opět vztažena k průměru, rozptylu, podílu hodnot i dalším statistickým parametrům popisujícím vzorek
- Kromě testů pro dvě skupiny hodnot existují samozřejmě i testy pro více skupin dat

One-tailed vs. Two-tailed testy

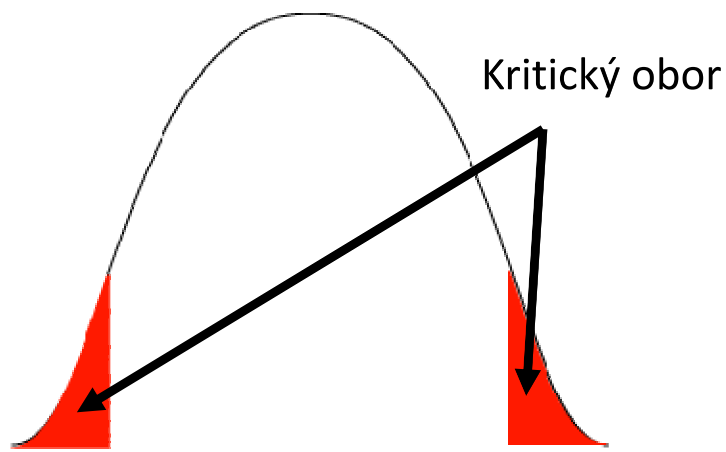
One – tailed testy

- Hypotéza testu je postavena asymetricky, tedy ptáme se na větší než/ menší než
- Test může mít pouze dvojí výstup – jedna z hodnot je větší (menší) než druhá a všechny ostatní případy



Two – tailed testy

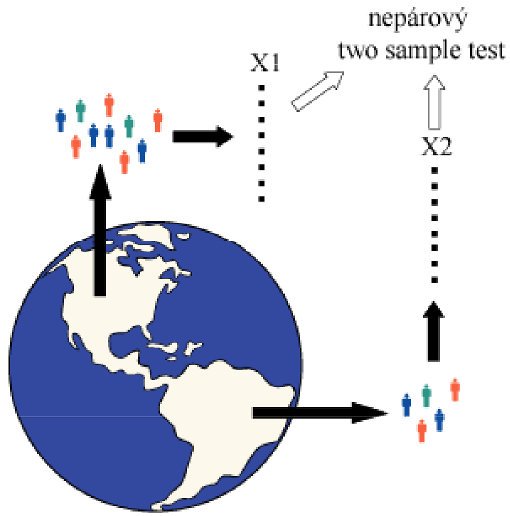
- Hypotéza testu se ptá na otázku rovná se/nerovná se
- Test může mít trojí výstup – menší - rovná se – větší než
- Situace nerovná se je tedy souhrnem dvou možných výstupů testu (menší+větší)



Nepárový vs. párový design

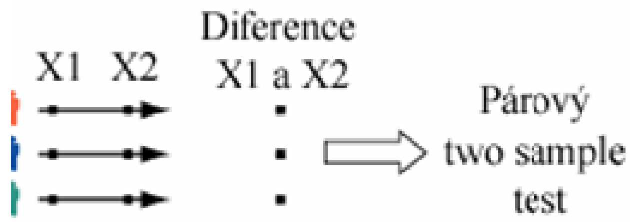
Nepárový design

- 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

Vícerozměrné statistické metody

Základní statistické testy

One sample t-test

V případě one sample testů jde o srovnání výběru dat (tedy one sample) s cílovou populací. Pro parametrické testy musí mít datový soubor normální rozložení.

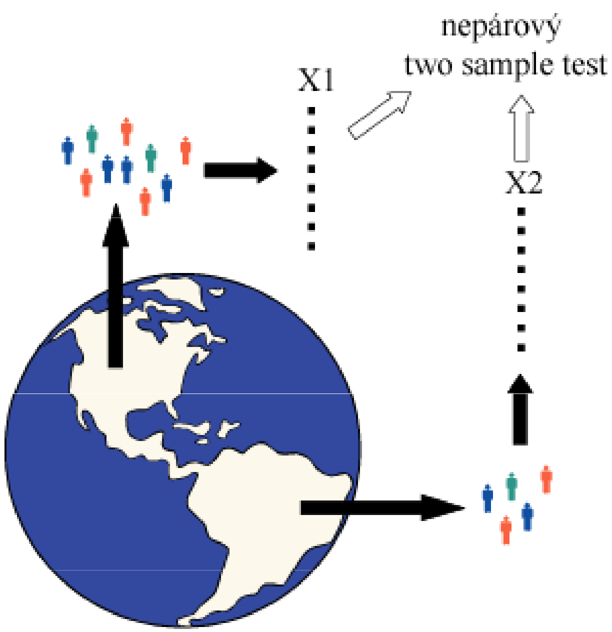
Průměr – cílová vs. výběrová populace

$$t = \frac{\bar{x} - \mu}{s}\sqrt{n}$$

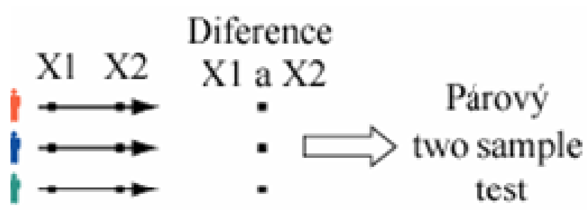
H_0	H_A	Testová statistika	Interval spolehlivosti
$\bar{x} \leq \mu$	$\bar{x} > \mu$	**t**	$t > t_{1-\alpha}^{(n-1)}$
$\bar{x} \geq \mu$	$\bar{x} < \mu$	**t**	$t < t_{\alpha}^{(n-1)}$
$\bar{x} = \mu$	$\bar{x} \neq \mu$	**t**	$\lvert t \rvert > t_{1-\alpha/2}^{(n-1)}$

Dvouvýběrové testy: párové a nepárové

- Při použití two sample testů srovnáváme spolu dvě rozložení. Jejich základním dělením je podle designu experimentu na testy párové a nepárové.



- Základním testem pro srovnání dvou nezávislých rozložení spojitých čísel je nepárový two-sample t-test



- Základním testem pro srovnání dvou závislých rozložení spojitých čísel je párový two-sample t-test

Dvouvýběrové testy: párové a nepárové

Data

X_1 X_2

Nezávislé uspořádání

X_1 X_2

$$H_0 : \mu_1 = \mu_2$$

n_1 n_2

$\bar{x}_2$ $\bar{x}_2$

s_1^2 s_2^2

Párové uspořádání

$X_1 - X_2 = D$

$$H_0 : \bar{D} = 0$$

n

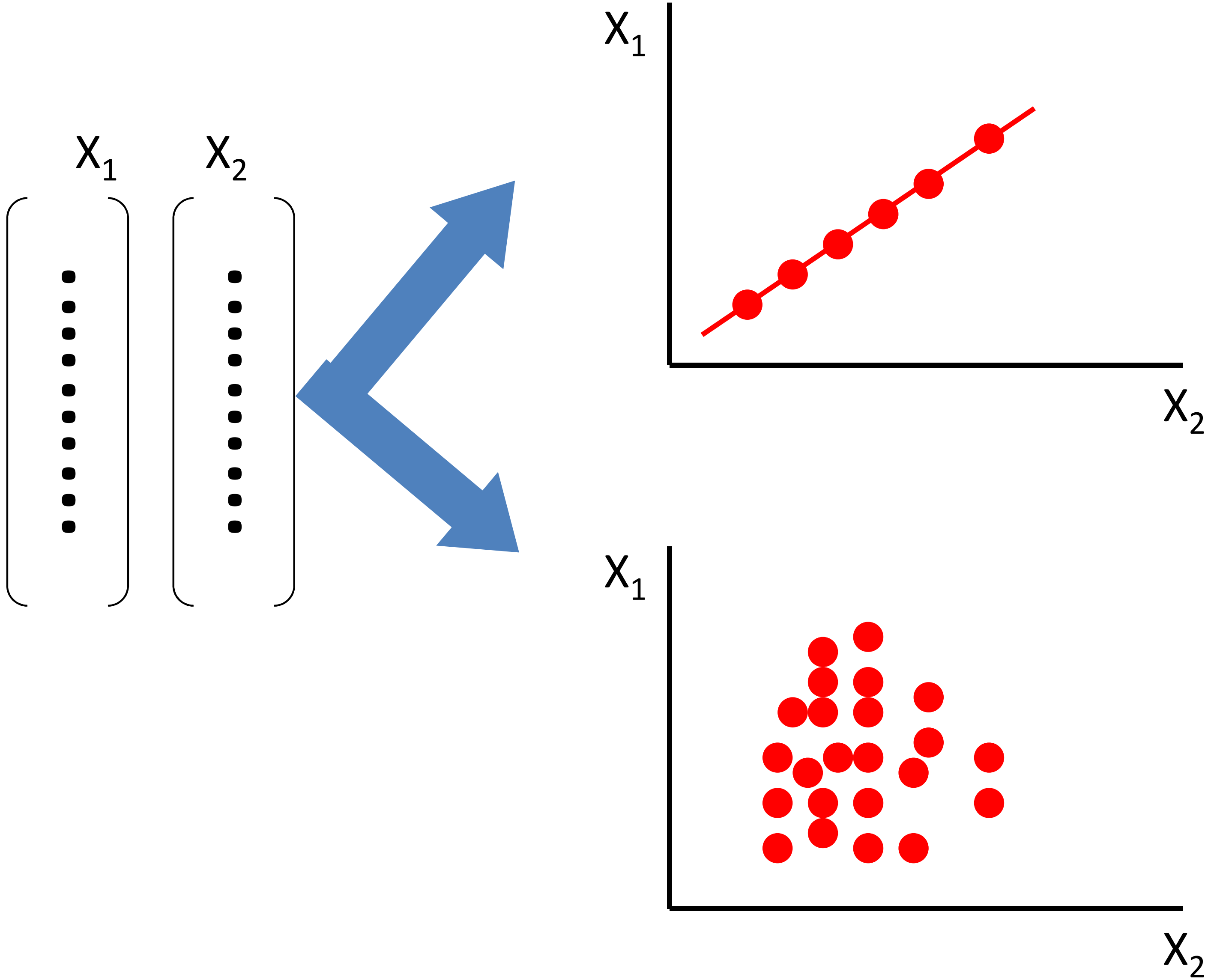
$\bar{D}$

s_D^2

$(n = n_2 = n_1)$

Design uspořádání
zásadně ovlivňuje interpretaci parametrů

Dvouvýběrové testy: párové a nepárové

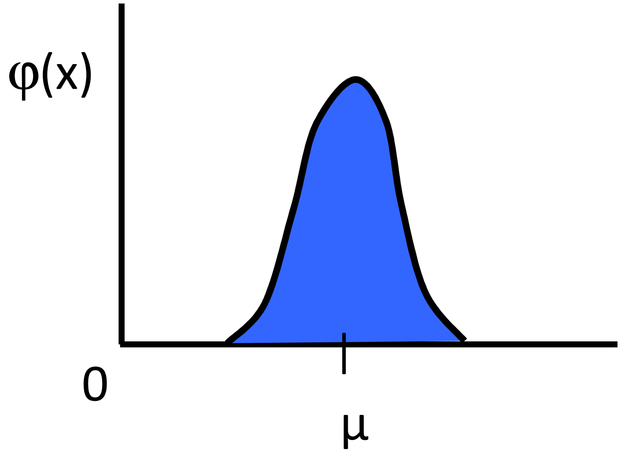


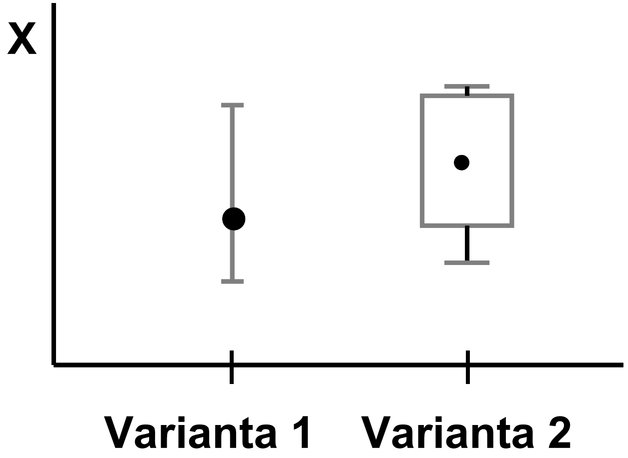
r = 0,954
(p < 0,001)

r = 0,218
(p < 0,812)

Předpoklady nepárového dvouvýběrového t-testu

- Náhodný výběr subjektů jednotlivých skupin z jejich cílových populací
- Nezávislost obou srovnávaných vzorků
- Přibližně normální rozložení proměnné ve vzorcích, drobné odchylky od normality ovšem nejsou kritické, test je robustní proti drobným odchylkám od tohoto předpokladu, normalita může být testována testy normality
- Rozptyl v obou vzorcích by měl být přibližně shodný (homoscedastic). Tento předpoklad je testován několika možnými testy – Levenův test nebo F-test.
- Vždy je vhodné prohlédnout histogramy proměnné v jednotlivých vzorcích pro okometrické srovnání a ověření předpokladů normality a homogenity rozptylu – nenahradí statistické testy, ale poskytne prvotní představu.





Nepárový dvouvýběrový t-test – výpočet I

1. nulová hypotéza: průměry obou skupin jsou shodné, alternativní hypotéza je, že nejsou shodné, two tailed test
2. prohlédnout průběh dat, průměr, medián apod. pro zjištění odchylek od normality a nehomogenita rozptylu, provést F –test

H_0	H_A	Testová statistika
$\sigma_1^2 \leq \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$	$F = \frac{s_1^2}{s_2^2}$
$\sigma_1^2 \geq \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$	$F = \frac{s_2^2}{s_1^2}$
$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	$F = \frac{\max(s_1^2; s_2^2)}{\min(s_1^2; s_2^2)}$

F-test pro srovnání dvou výběrových rozptylů

- Používá se pro srovnání rozptylu dvou skupin hodnot, často za účelem ověření homogenity rozptylu těchto skupin dat.

- V případě ověření homogenity je testována hypotéza shody rozptylů (two tailed); v případě shodných rozptylů je vše v pořádku a je možné pokračovat ve výpočtu t-testu, v opačném případě není vhodné test počítat.

Nepárový dvouvýběrový t-test – výpočet II

3. Výpočet testové statistiky (stupně volnosti jsou):

$$\upsilon = n_1 + n_2 - 2$$

$$t = \frac{Rozdíl\_průrůmě}{SE(rozdílprůo\ ěrů)} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$ vážený odhad rozptylu

4. výsledné t srovnáme s tabulární hodnotou t pro dané stupně volnosti a α (obvykle α=0,05)
5. Lze spočítat interval spolehlivosti pro rozdíl průměrů (např. 95%), počet stupňů volnosti a s^2 odpovídají předchozím vzorcům

$$(\bar{x}_1 - \bar{x}_2) \pm t_{0,975} SE(\bar{x}_1 - \bar{x}_2) = (\bar{x}_1 - \bar{x}_2) \pm t_{0,975}\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Dvouvýběrový t-test - příklad

Průměrná hmotnost ovcí v čase páření byla srovnávána pro kontrolní skupinu a skupinu krmenou zvýšenou dávkou potravy. Kontrolní skupina obsahuje 30 ovcí, skupina se zvýšeným příjmem potravy pak 24 ovcí.

- Vlastní experiment byl prováděn tak, že na začátku máme 54 ovcí (ideálně stejného plemene, stejně staré atd.), které náhodně rozdělíme do dvou skupin (náhodné rozdělování objektů do pokusných skupin je objektem celého specializovaného odvětví statistiky nazývaného randomizace). Poté co experiment proběhne, musíme nejprve ověřit teoretický předpoklad pro využití nepárového t-testu. Pro obě proměnné jsou vykresleny grafy (můžeme též spočítat základní popisnou statistiku), na kterých můžeme posoudit normalitu a homogenitu rozptylu, kromě okometrického pohledu můžeme pro ověření normality použít testy normality, pro ověření homogenity rozptylu pak F-test
- Pokud platí všechny předpoklady Two sample nepárového t-testu, můžeme spočítat testovou charakteristiku, výsledné t je 2,43 s 52 stupni volnosti, podle tabulek je a $t_{0,975\,(52)}$= 2,01, tedy t> $t_{0,975\,(52)}$= a nulovou hypotézu můžeme zamítnout, skutečná pravděpodobnost je pak 0,018. Rozdíl mezi skupinami je 1,59 kg ve prospěch skupiny s lepší výživou.

$$t = \frac{Rozdíl\_průrůmě}{SE(rozdílprůo\ ěrů)} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \qquad \upsilon = n_1 + n_2 - 2$$

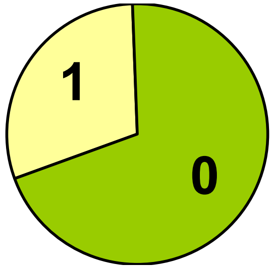
- Pro rozdíl mezi oběma soubory jsou spočítány 95% konfidenční intervaly jako 1,59±2.01*(0,655) kg, což odpovídá rozsahu 0,28 až 2,91 kg. To, že konfidenční interval nezahrnuje 0 je dalším potvrzením, že mezi skupinami je významný rozdíl – jde o další způsob testování významnosti rozdílů mezi skupinami dat – nulovou hypotézu o tom, že rozdíl průměrů dvou skupin dat je roven nějaké hodnotě zamítáme v případě, kdy 95% konfidenční interval rozdílu nezahrnuje tuto hodnotu (v tomto případě 0).

$$(\bar{x}_1 - \bar{x}_2) \pm t_{0,975} SE(\bar{x}_1 - \bar{x}_2) = (\bar{x}_1 - \bar{x}_2) \pm t_{0,975}\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Test dobré shody - základní teorie

Binomické jevy (1/0)

$$\chi^2_{(1)} = \underbrace{\frac{\left[\text{pozorovaná četnost} - \text{očekávaná četnost}\right]^2}{\text{očekávaná četnost}}}_{\text{I. jev 1}} + \underbrace{\frac{\left[\text{pozorovaná četnost} - \text{očekávaná četnost}\right]^2}{\text{očekávaná četnost}}}_{\text{II. jev 2}}$$



Příklad

✓ 10 000 lidí hází mincí → rub: 4 000 případů (R)
→ líc: 6 000 případů (L)

? Lze výsledek považovat za statisticky významně odlišný (nebo neodlišný) od očekávaného poměru R : L = 1 : 1 ?

$$\boxed{\chi^2_{(1)} = \frac{(4000 - 5000)^2}{5000} + \frac{(6000 - 5000)^2}{5000} = 400}$$

Tabulková hodnota: $\chi^2_{(0,95)}(\nu = 1) = \underline{\underline{3,84}}$ $(0,95 = 1 - \alpha)$

Rozdíl je vysoce statisticky významný (p << 0,001]

Kontingenční tabulky - 0 :Nezávislost dvou jevů A a B

Kontingenční tabulka 2 x 2

B ↓ \ A →	+	-	Podíl (+)
+	a	b	$\frac{a}{(a+b)}$ p_1
-	c	d	$\frac{c}{(c+d)}$ p_2
Podíl (+)	$\frac{a}{(a+c)}$	$\frac{b}{(b+d)}$	

N = a + b + c + d

$$P(B^+) = \frac{(a+b)}{N}$$

$$P(B^-) = \frac{(c+d)}{N}$$

Očekávané četnosti:

$$F_{(A)} = \frac{(a+b)(a+c)}{N} \qquad F_{(C)} = \frac{(a+c)(d+c)}{N}$$

$$F_{(B)} = \frac{(a+b)(b+d)}{N} \qquad F_{(D)} = \frac{(b+d)(c+d)}{N}$$

$$\nu = 1 = (r-1)*(c-1)$$

$$P_{(A)}; P_{(B)}$$

$$\chi^2_{\nu=1} = \sum_{i=1}^{4} \frac{(f_i - F_i)^2}{F_i}$$

$$\chi^2_c = \sum \sum \frac{\left(\left|f_{ij} - F_{ij}\right| - 0,5\right)^2}{F_{ij}}$$

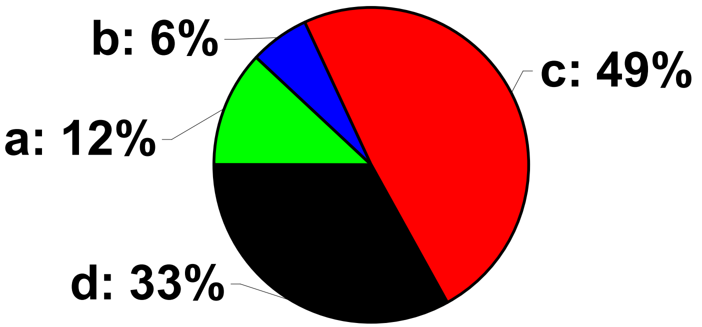
Kontingenční tabulky: příklad

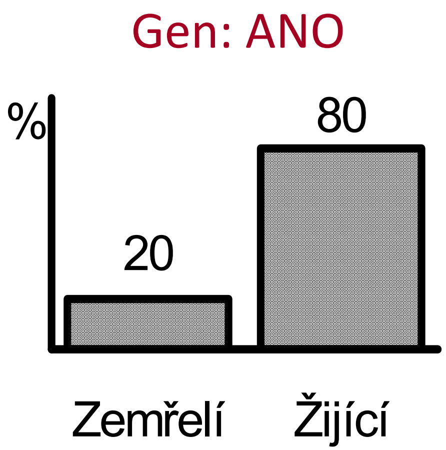
gen \ †	**Ano**	**Ne**	Σ
Ano	20	82	102
Ne	10	54	64
Σ	30	136	166

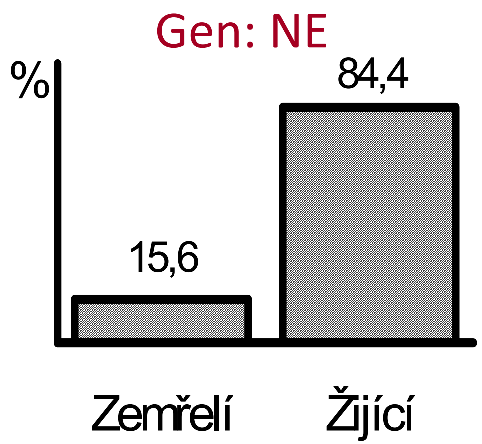
F_A = 102 * 30 / 166 = 18,43
F_B = 102 * 136 / 166 = 83,57
F_C = 11,57
F_D = 52,43

$$\chi^2_{(1)} = \frac{(20-18{,}43)^2}{18{,}43} + \frac{(82-83{,}57)^2}{83{,}57} + \frac{(10-11{,}57)^2}{11{,}57} + \frac{(54-52{,}43)^2}{52{,}43} = 0{,}423 \qquad 0{,}423 < \chi^{2\,(1)}_{0,95} = 3{,}84$$

Kontingenční tabulka v obrázku







ANOVA – základní výpočet

- Základním principem ANOVY je porovnání rozptylu připadajícího na:
  - Rozdělení dat do skupin (tzv. effect, variance between groups)
  - Variabilitu objektů uvnitř skupin (tzv. error, variance within groups), předpokládá se, že jde o náhodnou variabilitu (=error)

1. Variabilita mezi skupinami

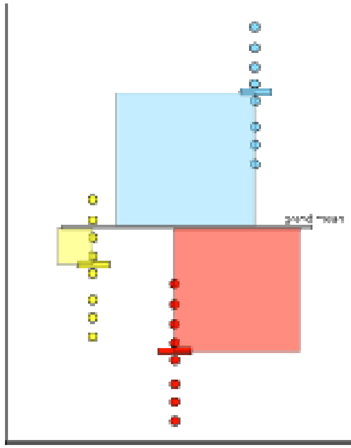
Rozptyl je počítán pro celkový průměr (tzv. grand mean) a průměry v jednotlivých skupinách dat

Stupně volnosti jsou odvozeny od počtu skupin (= počet skupin -1)

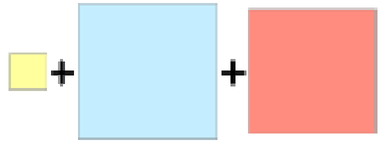
2. Variabilita uvnitř skupin

Rozptyl je počítán pro průměry jednotlivých skupin a objekty uvnitř příslušných, celková variabilita je pak sečtena pro všechny skupiny

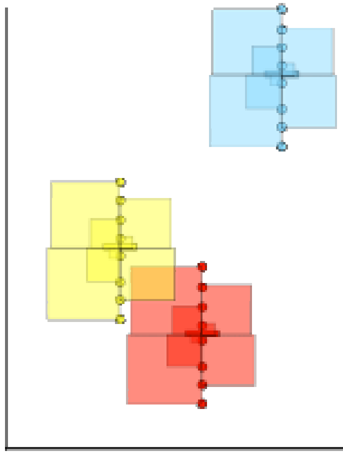
Stupně volnosti jsou odvozeny od počtu hodnot (= počet hodnot - počet skupin)



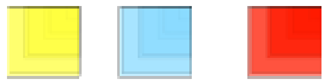
SS between



$$\nu_1 = k - 1$$



SS within



$$\nu_2 = n - k$$

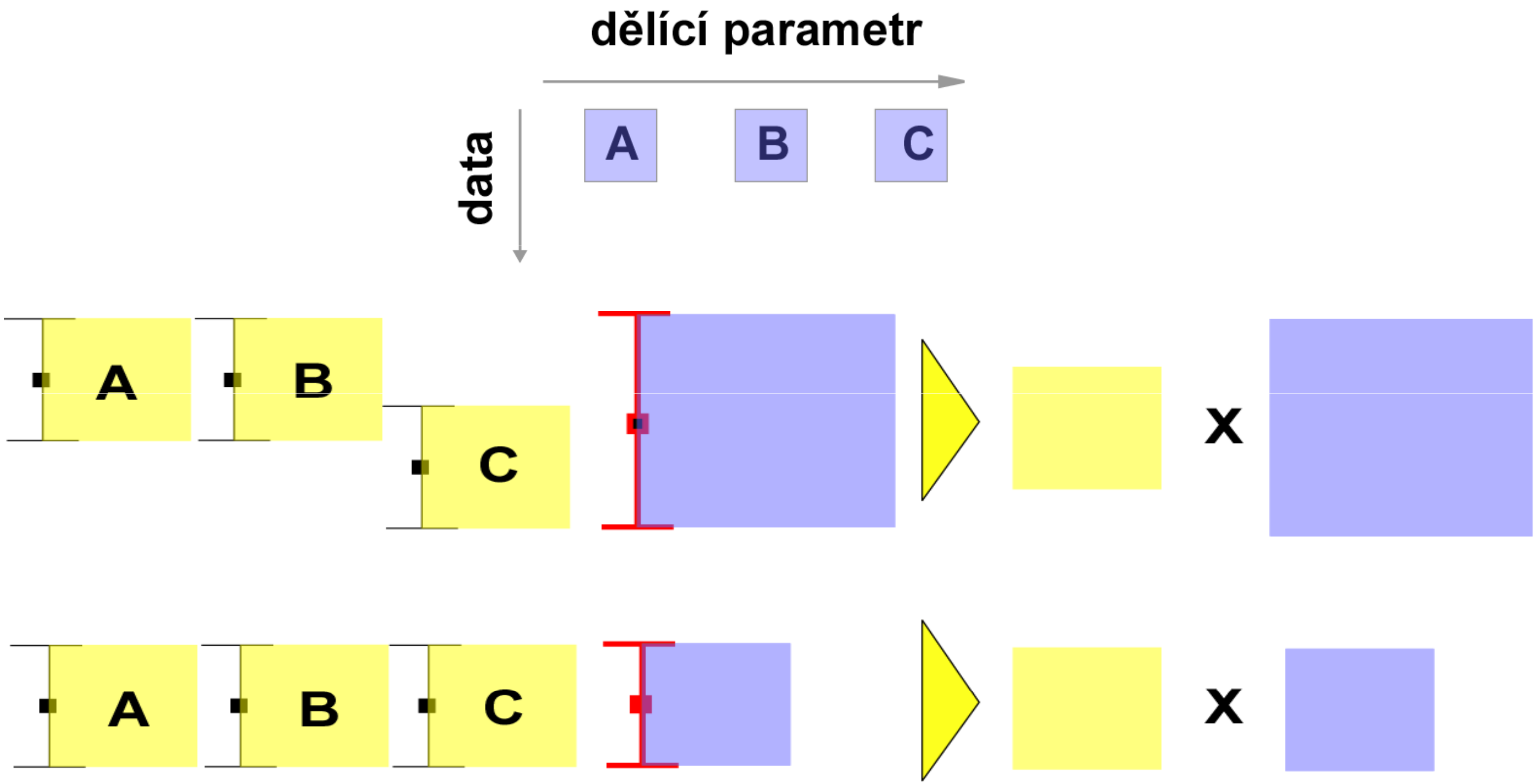
$$F = \frac{between\_groups}{within\_groups}$$

<u>Výsledný poměr (F) porovnáme s tabulkami F rozložení pro v_1 a v_2 stupňů volnosti</u>

SS=sum of squares

Jednoduchý ANOVA design

Nejjednodušším případem ANOVA designu je rozdělení na skupiny podle jednoho parametru.



Nested ANOVA

- Rozdělení skupin na náhodné podskupiny (např. opakování experimentu)
- Cílem je zjistit, zda data v jedné skupině nejsou pouhou náhodou
- Nejprve je testována shoda podskupin v hlavních skupinách,
  - pokud jsou shodné, je vše v pořádku
  - pokud nejsou, stále lze zjišťovat, zda se variabilita uvnitř hlavních skupin liší od celkové variability

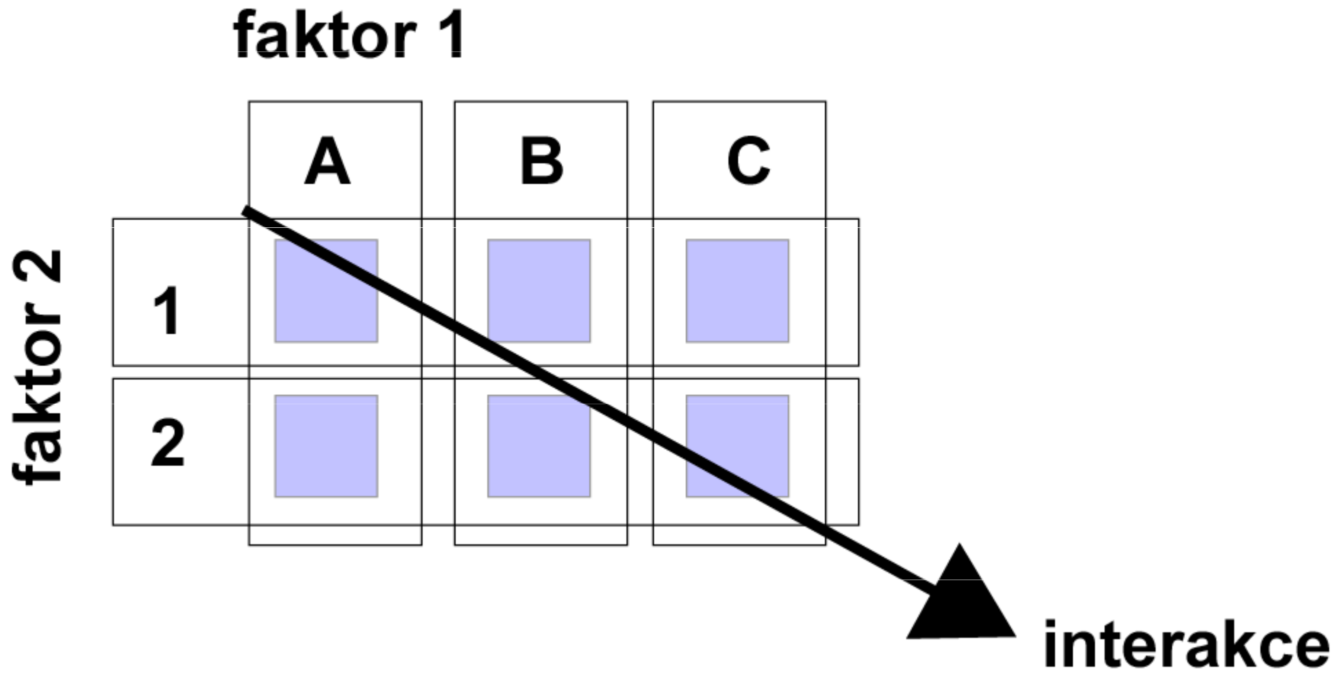
jednoduchá ANOVA **nested ANOVA**

Two way ANOVA

Pro rozdělení do kategorií je zde více parametrů

Na rozdíl od nested ANOVY nejde o náhodná opakování experimentu, ale o řízené zásahy (např.vliv pH a koncentrace O_2)

Kromě vlivu hlavních faktorů se uplatňuje i jejich interakce



Modely analýzy rozptylu - základní výstup

Základním výstupem analýzy rozptylu je
Tabulka ANOVA - frakcionace komponent rozptylu

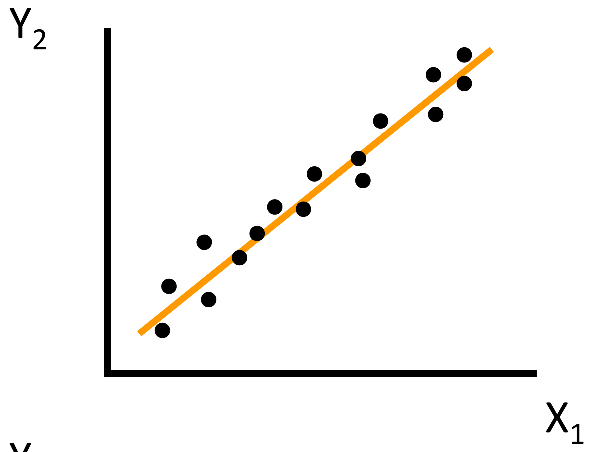
Zdroj rozptylu	St. v.	SS	MS	F
Pok. zásah (mezi skupinami)	a -1	SS_B	$SS_B/(a-1)$	MS_B/MS_E
Uvnitř skupin	N - a	SS_E	$SS_E/(N-a)$	
Celkem	N -1	SS_T		

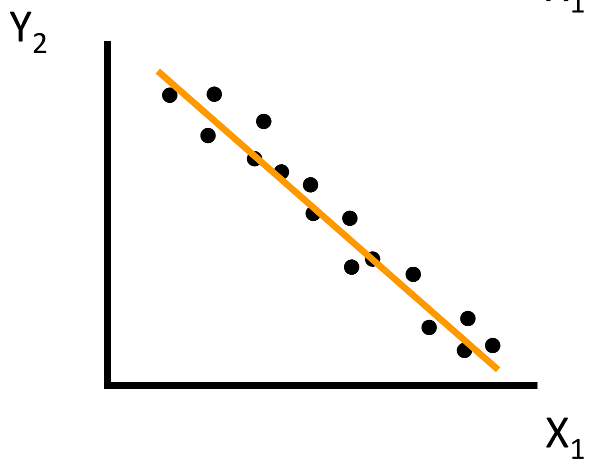
SS_B/SS_T → Kvantifikovaný podíl rozdílu mezi pokusnými zásahy na celkovém rozptylu

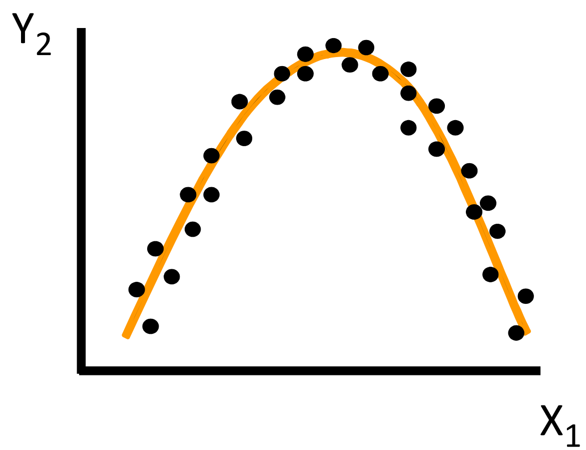
MS_B/MS_T → Statistická významnost rozdílu

Základy korelační analýzy I

Korelace - vztah (závislost) dvou znaků (parametrů)







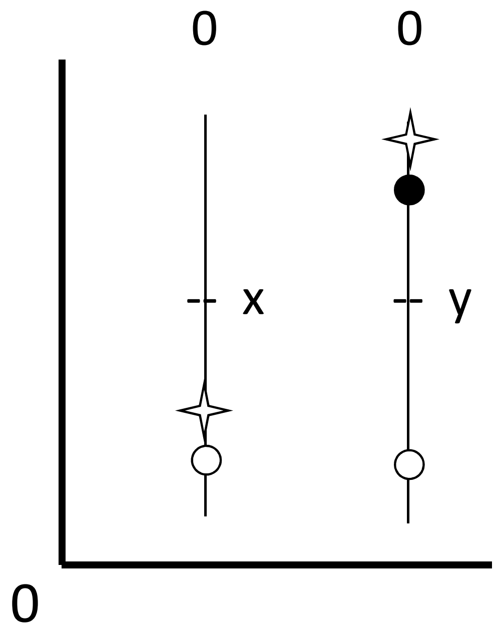
X_2 \ X_1	**ANO**	**NE**
ANO	a	b
NE	c	d

Základy korelační analýzy II

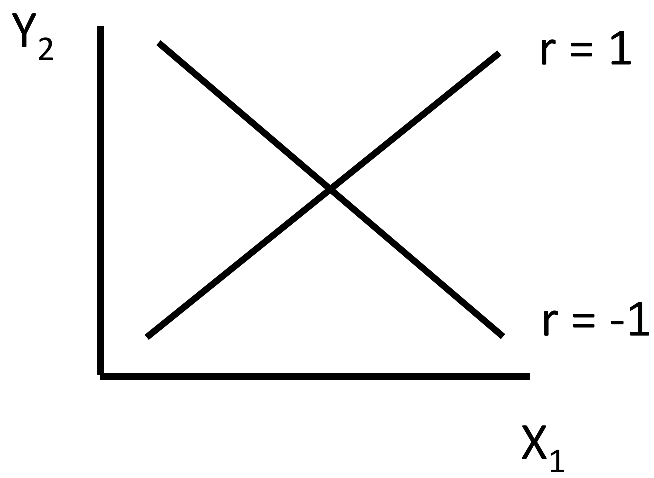
Parametrické míry korelace

Kovariance

$$Cov(x,y) = E(x_i - \bar{x}).(y_i - \bar{y})$$



Pearsonův koeficient korelace



Základy korelační analýzy III

P_I (zem)	10	14	15	32	40	20	16	50
P_I (rostl.)	19	22	26	41	35	32	25	40

$I = 1, \ldots, n; n = 8; v = 6$

$$r = \frac{Cov(x,y)}{S_x . S_y} = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sqrt{\left[\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2\right]\left[\sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2\right]}} = 0{,}7176$$

I. $H_0 : \rho = \phi : \alpha = 0{,}05$

tab : $r(v = 6) = 0{,}7076$

II. $H_0 : \rho = \phi$

$$t = \left[\frac{r}{\sqrt{1 - r^2}}\right] \cdot \sqrt{n - 2}$$

$v = n - 2$

$$\left.\begin{array}{l} t = \dfrac{0{,}7176}{0{,}6965} \cdot \sqrt{6} = 2{,}524 \\ \textbf{tab} : t_{0{,}975}^{(n-2)} = 2{,}447 \end{array}\right\}$$

$P \leq 0{,}05$

Základy regresní analýzy

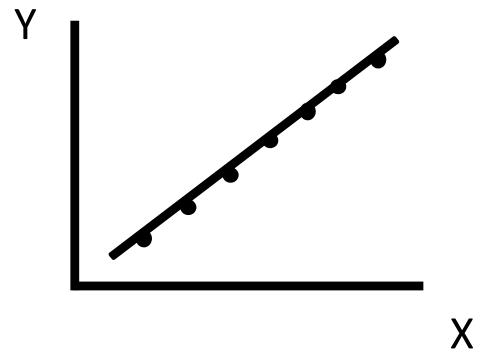
Regrese - funkční vztah dvou nebo více proměnných

Jednorozměrná
$y = f(x)$

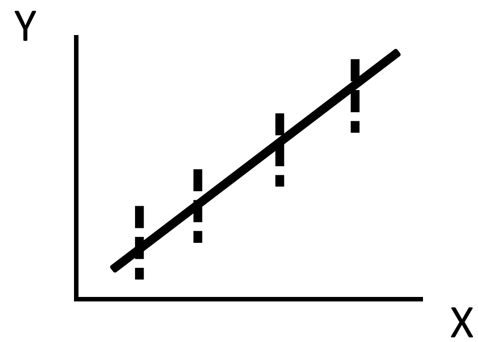
Vícerozměrná
$y = f(x1, x2, x3, \ldots\ldots xp)$

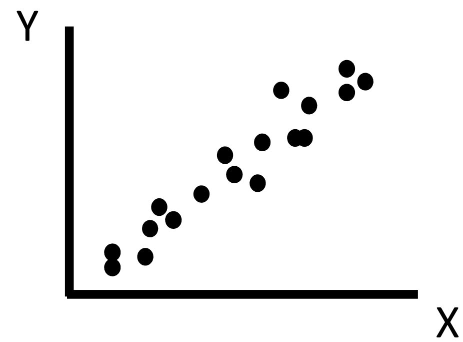
Vztah x, y

Deterministický



Regresní, stochastický





Pro každé x existuje pravděpodobnostní rozložení y

Regresní analýza přímky: lineární regrese

$$Y = a + b \cdot x + e \quad \approx \quad \alpha + \beta \cdot X + \varepsilon$$

y

$\alpha \approx a$ **(intercept):** $a = \overline{y} - b \cdot \overline{x}$

$\beta \cdot X \approx b \cdot x$ **(sklon; slope)**

$\varepsilon \approx e$ **- náhodná složka:** $\mathrm{N}(0; \sigma_e^2) = N(0; \sigma_y^2 x)$

Komponenty tvořící y se sčítají

ε - náhodná složka modelu přímky = rezidua přímky

$\sigma_e^2 (\sigma_{y \cdot x}^2) \Rightarrow$ **rozptyl reziduí**