Úvod do statistiky

B02907 Informační a komunikační technologie

Lubomír Štěpánek[1, 2]

[1]Oddělení biomedicínské statistiky & výpočetní techniky
Ústav biofyziky a informatiky
1. lékařská fakulta
Univerzita Karlova v Praze

ČVUT FBMI

[2]Katedra biomedicínské informatiky
Fakulta biomedicínského inženýrství
České vysoké učení technické v Praze

22. února 2018

Obsah

Co je statistika

- statistika zkoumá jevy, které se projeví až na velkém souboru případů, nikoliv pouze v jednom případě
- statistika shromažďuje, zpracovává a kvantitativně interpretuje data
- původní význam pojmu statistika souvisí se státem – správa, daně, výměry

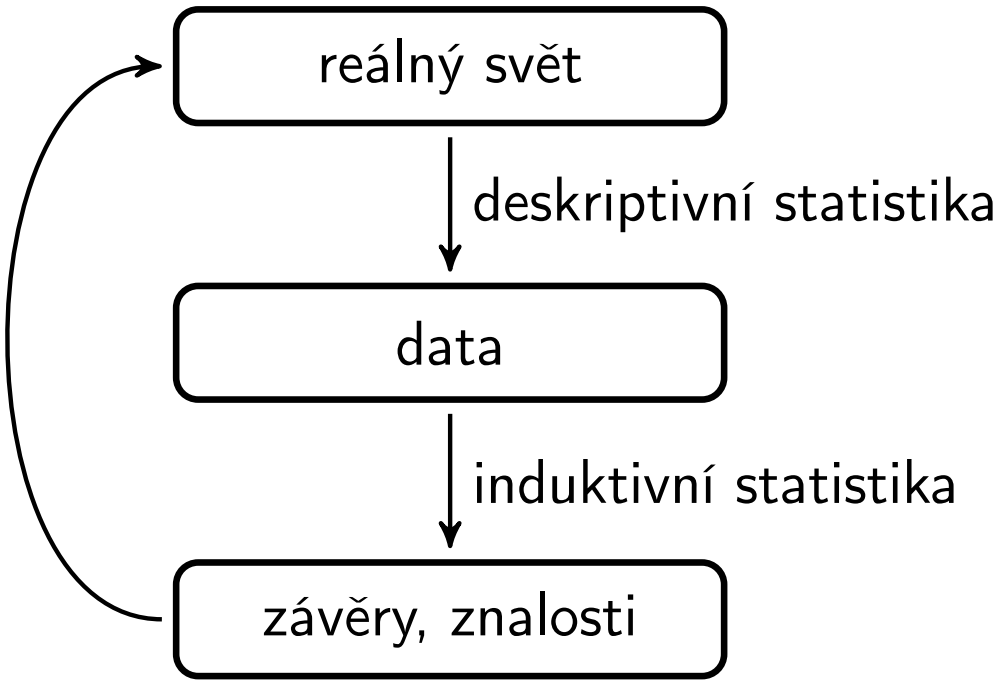
Role statistiky ve výzkumu a závěrečné práci

- v biomedicíně je statistické zpracování dat a následná interpretace mnohdy hlavní kvantitativní složkou výzkumné práce
- řídí se standardy a dobrou praxí, umožňuje závěry hodnotit a porovnávat
- pro daná vstupní data a hypotézy[1] obvykle existuje postup, který je považován za nejlepší možný, nebo alespoň vhodný
- je vhodné statistické zpracování dat v závěrečné práci nepodcenit

[1] alespoň na úrovni závěrečných prací

Dělení statistiky

- deskriptivní statistika
  - popisuje data, ale nedělá na nich žádné „velké“ závěry
- induktivní statistika
  - pozoruje konkrétní data a vyvozuje z nich obecné závěry, ovšem s udáním stupně jejich spolehlivosti



Základní pojmy

- statistický znak, veličina
  - měřitelná (veličina) či jinak zjistitelná (znak) charakteristika našeho zájmu
  - např. tělesná výška, pohlaví, mzda, apod.
- statistická jednotka
  - základní atomický prvek zájmu, u nějž lze měřit nebo jinak získat hodnotu statistického znaku či veličiny
  - např. student, pacient, stát, molekula, apod.
- statistický soubor
  - množina statistických jednotek (prvků statistického souboru)
  - např. třída žáků, kohorta pacientů, apod.

Pojem *populace*

- **populace := základní soubor**
- úplná množina (statistický soubor) všech prvků (statistických jednotek), které spojuje určitá vlastnost a o kterých se snažíme statisticky něco zjistit
- prvky dány výčtem (je-li rozsah populace konečný), nebo společnou vlastností všech prvků (je-li rozsah populace nekonečný i konečný)
- rozsah konečně velké populace obvykle značíme N (u nekonečně velké populace $N \to \infty$)
- např. {T. G. Masaryk, E. Beneš, . . . , V. Klaus, M. Zeman}, {všichni dosavadní prezidenti českého státu}, {všichni obyvatelé Evropy}, apod.

Pojem *výběr*

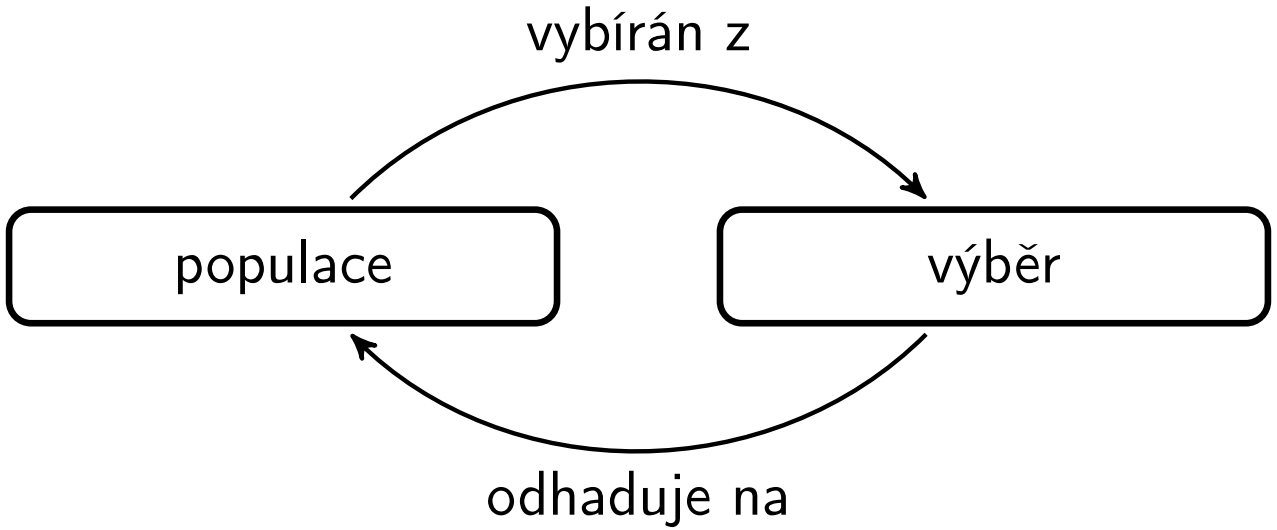
- vyšetřit celou populaci v praxi takřka nemožné
- nekonečně velké populace nelze celkově šetřit už z principu
- výběr := statistický soubor, obsahuje vybrané prvky z populace; je tedy podmnožinou populace
- výběr pořizujeme metodou náhodného, či záměrného výběru
- cílem získat reprezentativní výběr (vystihuje vlastnosti populace), nikoliv selektivní výběr

Reprezentativní výběr

- takový výběr, z kterého je induktivními metodami možné usuzovat na vlastnosti „mateřské“ populace
- pořizujeme *záměrným*, či *náhodným* výběrem
  - záměrný výběr – opírá se o expertízu, zatížen subjektivitou
  - náhodný výběr – náhodné, nezávislé vybírání prvků populace do výběru

Vztah populace a výběru

- z populace je vybírán výběr
- z charakteristik výběru jsou odhadovány charakteristiky populace



Statistické znaky a veličiny

- každý prvek statistického souboru má svou hodnotu[2] určitého zkoumaného statistického znaku či veličiny (jde-li o měřitelný znak)
- např. *ve školní třídě změříme tělesnou výšku každého žáka*
  - *školní třída* je statistický soubor
  - *žáci* jsou statistické prvky (jednotky)
  - *tělesná výška* je statistická veličina

[2] ta může eventuálně chybět nebo být neznámá (missing value)

Kvantitativní znak (veličina)

- je vyjádřen číslem (a obvykle s jednotkou), kdy s číselnou hodnotou je smysluplné provádět aritmetické operace
- číslo tedy nenese pouze „katalogizační“ význam
- někdy též označován jako *metrický* typ dat
- dle spojitosti číselných hodnot
  - *spojitý* – hodnoty nabývají reálných čísel, nebo je na ně lze převést nějakou bijekcí; např. hmotnost, výška atd.
  - *diskrétní* – hodnoty jsou oddělená čísla obvykle ve smyslu počet či pořadí, např. počty pacientů atd.
- dle měřítka
  - *intervalová stupnice* – lze si smysluplně odpovědět, o kolik se dvě hodnoty liší, ale ne kolikrát; např. °C, datumy atd.
  - *poměrová stupnice* – lze si smysluplně odpovědět, o kolik se dvě hodnoty liší i kolikrát se liší; např. °K

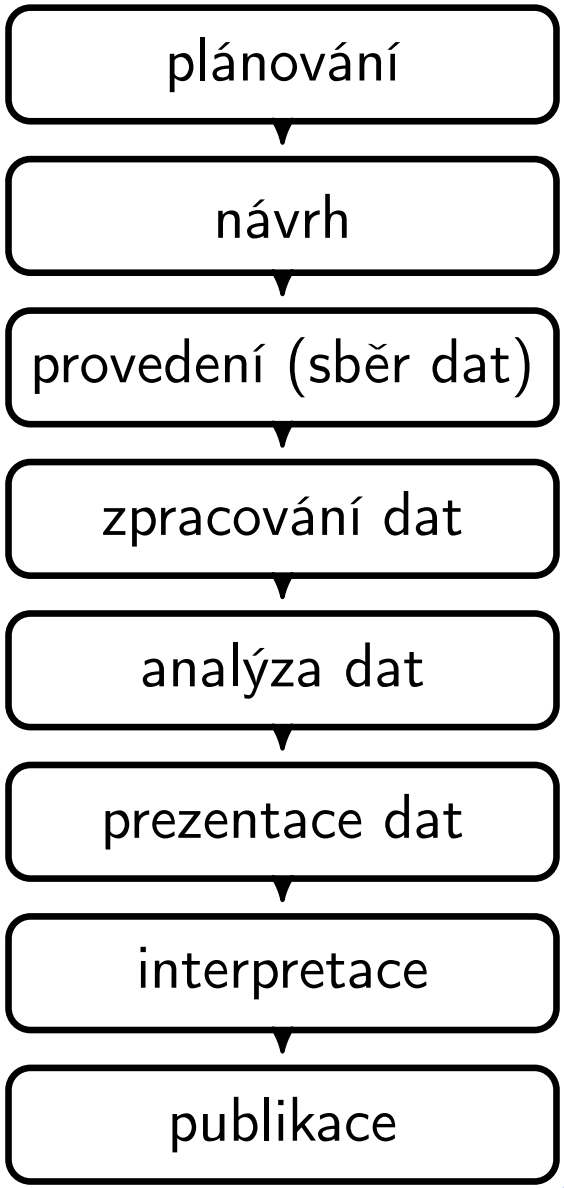
Kvalitativní znak

- je vyjádřen obvykle slovně
- pokud číslem, pak nese pouze „katalogizační“ význam a není smysluplné s ním provádět aritmetické operace
- někdy též označován jako *ordinální* či *alternativní* typ dat
- dle měřítka
  - *nominální stupnice* – dvě či více vzájemně se vylučujících, rovnocenných tříd, které nelze uspořádat na číselné ose; např. pohlaví {muž, žena}, rodinný stav muže {svobodný, ženatý, rozvedený, vdovec, registrovaný}
  - *ordinální stupnice* – kategorie je možné uspořádat vzestupně/sestupně, lze si smysluplně odpovědět, která hodnota je větší než jiná (ale ne o kolik, natož kolikrát); např. pořadí v závodu, grade tumoru {1, 2, 3, 4} atd.

Intermezzo

- určete typ znaku a stupnice u následujících příkladů
  - procentuální úspěšnost v testu v souboru studentů jednoho kruhu [%]
  - soubor všech červencových dní jednoho roku (1., 2., . . . , 31.)
  - soubor čísel všech autobusů projíždějících zastávkou Kajetánka (174, 180, . . .)
  - soubor mutací genu CFTR (F508del, . . .)
  - bolest hodnocená pomocí VAS [0–10]
  - počet porodů v jedné porodnici za jednu noc
  - staging kolorektálního karcinomu {1, 2, 3, 4}

Obecné schéma výzkumné studia & uplatnění statistiky



Plánování a návrh studie

- přesná formulace cíle a účelu výzkumu
  - např. „poznat účinky dvou antidiabetik a používat lepší z nich“
  - vhodná spolupráce se statistikem, formulace hypotéz (ne až na závěr!)
- vymezení pojmů a metod
  - studovaná populace (šetření úplné vs. výběrové; metoda výběru, odhad rozsahu dat)
  - sledované znaky (přesná definice, povaha znaku, stupnice měření)
  - sběr dat (observace, rozhovor, dotazník, dokumentace)
  - statistická analýza, technické zpracování dat (ručně vs. počítačově)

Sběr a zpracování dat studie

- pilotní studie – ověření metod pozorování a měření na malém vzorku populace
- vyřazení neadekvátních prvků výběru (prvky s nekvalitní dokumentací, nevyplněný dotazník)
- vyloučení formálních chyb (např. „prvek narozen roku 2135“)
- snaha vyvarovat se náhodným (z nepozornosti) a systematickým chybám (špatná kalibrace, nevhodné otázky v dotazníku)
- dnes zpracování dat obvykle počítačem

Analýza výsledků studie

- statistické třídění
  - rozdělení prvků souboru do skupin podle společných vlastností znaku
  - kvalitativní znak — např. znak pohlaví do dvou tříd {muž, žena}
  - kvantitativní znak — třídy tvoří sousedící nepřekrývající se intervaly; např. znak věk pacientů do tříd [0-10], [10-20], [20-30], [30-40], [40-50], [50-60], [60-70], [70+]
- určení četností tříd
  - *absolutní četnost* — počet prvků dané třídy
  - *relativní četnost* – absolutní četnost dělená rozsahem výběru
- určení, výpočet ukazatelů
  - deskriptivní a induktivní statistika – pravděpodobnostní rozdělení (normalita dat), odlehlé hodnoty, střední hodnota a variabilita výběru a populace, stanovení chyb alfa a beta, hladiny významnosti atd.

Interpretace a prezentace výsledků studie

- interpretace výsledků
  - vhodná konzultace se statistikem
- prezentace výsledků
  - slovním výkladem, tabulkou, diagramem

Publikace výsledků studie

- snaha zajistit reprodukovatelnost a kvalitu
- nekvalitní zpracování – nepohodlí, zneužívání času pacientů a výzkumníků, zbytečná statistická šetření
- nekvalitní závěry – hrozba chybné klinické aplikace výzkumu
  - ***chyba prvního typu*** – potvrdíme fakt, který ve skutečnosti neplatí; např. „lék A funguje jako kardiostimulans, kardiak ho užívá a zemře na zástavu srdce“
  - ***chyba druhého typu*** – zamítneme fakt, který ve skutečnosti platí; např. „lék B se zkoušel v léčbě nádorů, avšak výzkum nepotvrdil jeho efekt; takže lék B propadl dějinám (přestože by fungoval!) a k pacientům se nemohla dostat fungující léčba“

Publikace výsledků studie

- nutné kriticky kontrolovat vlastní práci
  - někdy snaha publikovat za každou cenu („publish or perish!“)
- kontrolovat práci ostatních
  - protekce, neznalost „odborníků“, profitování ze spoluautorství
- kriticky nastavit vydavatelskou politiku
  - touha po senzaci (časopisy raději otiskují pozitivní závěry (lék funguje!) než negativní (lék nefunguje!), přestože jsou oba výsledky mnohdy stejně přínosné („věda nezná neúspěch“)

Literatura


Karel Zvára. *Biostatistika*. Praha: Karolinum, 2003. ISBN: 978-80-246-0739-9.


Jana Zvárová. *Základy statistiky pro biomedicínské obory*. Praha: Karolinum, 2016. ISBN: 978-80-246-3416-6.

Děkuji za pozornost!

lubomir.stepanek@lf1.cuni.cz
lubomir.stepanek@fbmi.cvut.cz