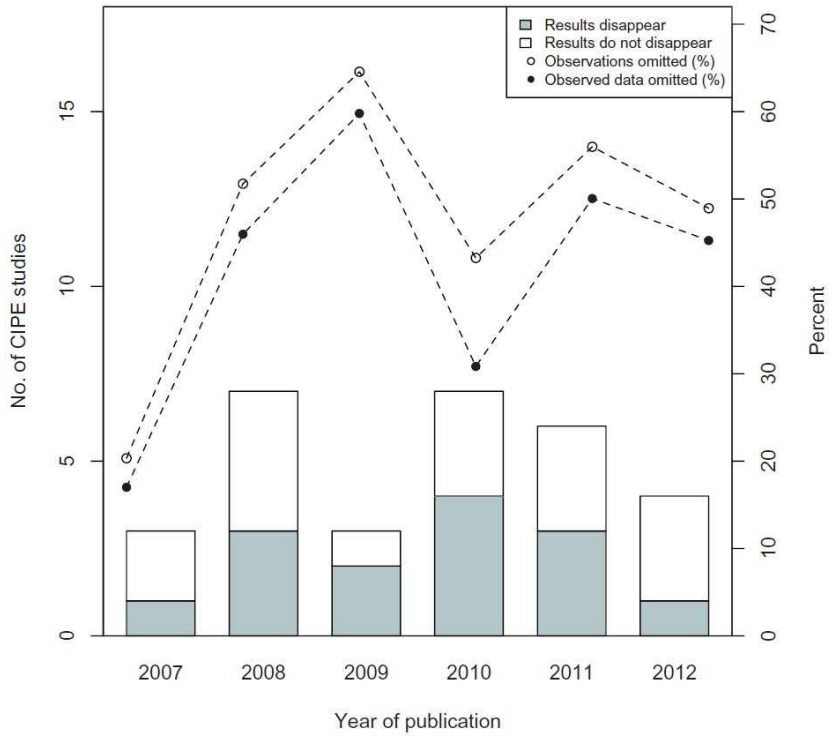
Chybějící data v praxi: jak se vyhnout neštěstí způsobenému listwise deletion?

Ivan Petrúšek
ivan.petrusek@soc.cas.cz
3. 2. 2017

Chybějící data

- „prázdná“ místa v datové matici (tzn. u některých proměnných a některých případů **nejsou hodnoty pozorovány)**
- **Předpoklad:** chybějící hodnoty **„zakrývají“ skutečné hodnoty**, které by jinak byly smysluplnou součástí analýzy
- Standardní **statistické metody byly vyvinuty pro data bez chybějících hodnot**. Výzkumník vždy musí zvolit strategii, jak s chybějícími hodnotami pracovat:
  - **Listwise deletion**
  - **Mnohonásobné imputace**

Jsou publikované výsledky zpochybnitelné kvůli používání listwise deletion?



Zdroj: R. Lall: *How Multiple Imputation Makes a Difference* [Political Analysis 2016, 24 (4), pp. 1–20]

Listwise Deletion

- **Analýza kompletních případů**
- **Každý případ, u kterého chybí alespoň jedna hodnota** (u některé z proměnných vstupujících do analýzy), **je z analýzy vyřazen**
- V programech na analýzu dat se jedná **většinou o defaultní metodu** práce s chybějícími hodnotami

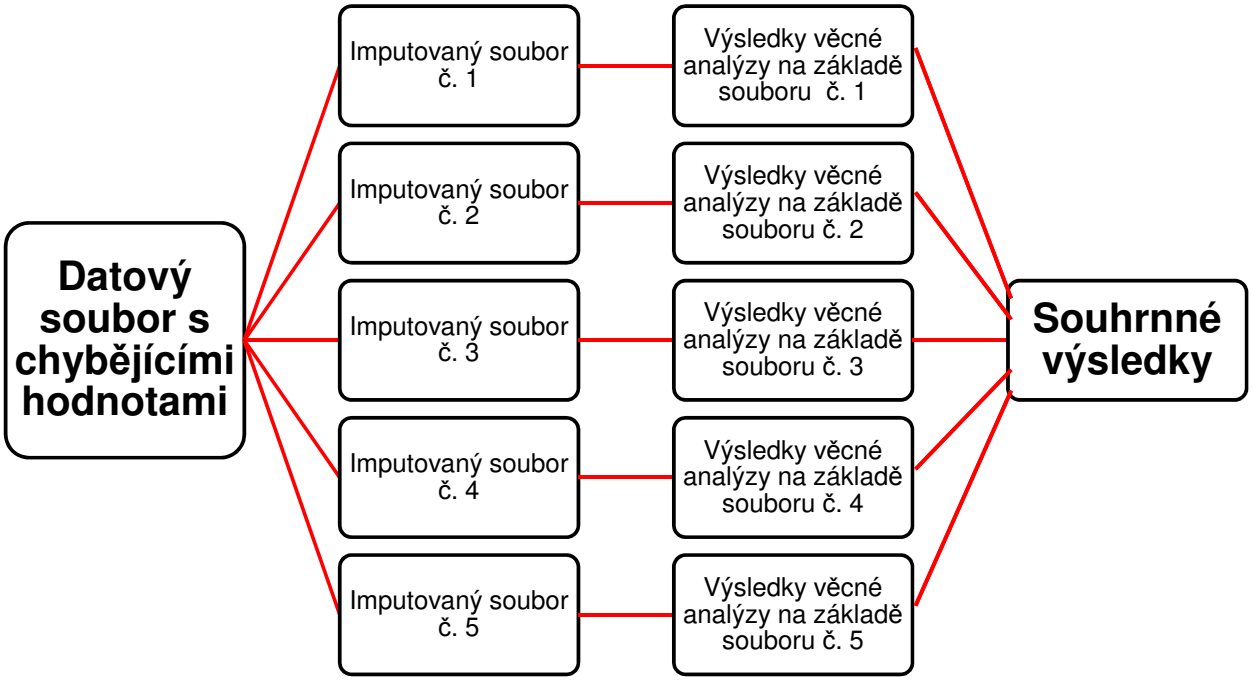
Listwise Deletion

- ✓ Snadné provedení
- ✗ Údaje pozorované u zbylých proměnných jsou v analýze zcela nevyužity
- ✗ Metoda vede k **nevychýleným odhadům** zkoumaných parametrů **pouze když hodnoty chybí zcela náhodně**
- ✗ **Snižuje počet pozorování** (n) → **větší hodnoty směrodatných chyb odhadů** (delší intervaly spolehlivosti a vyšší *p* hodnoty)

Mnohonásobné imputace

- **Každá chybějící hodnota je současně nahrazena několika verzemi**
- Rozptýlenost jednotlivých hodnot napříč nahrazenými verzemi má zohlednit stupeň nejistoty spojený s procesem nahrazování
- Jedná se o komplexní přístup, v rámci kterého existuje množství různých algoritmů, jak chybějící hodnoty nahrazovat

Schéma postupu mnohonásobných imputací



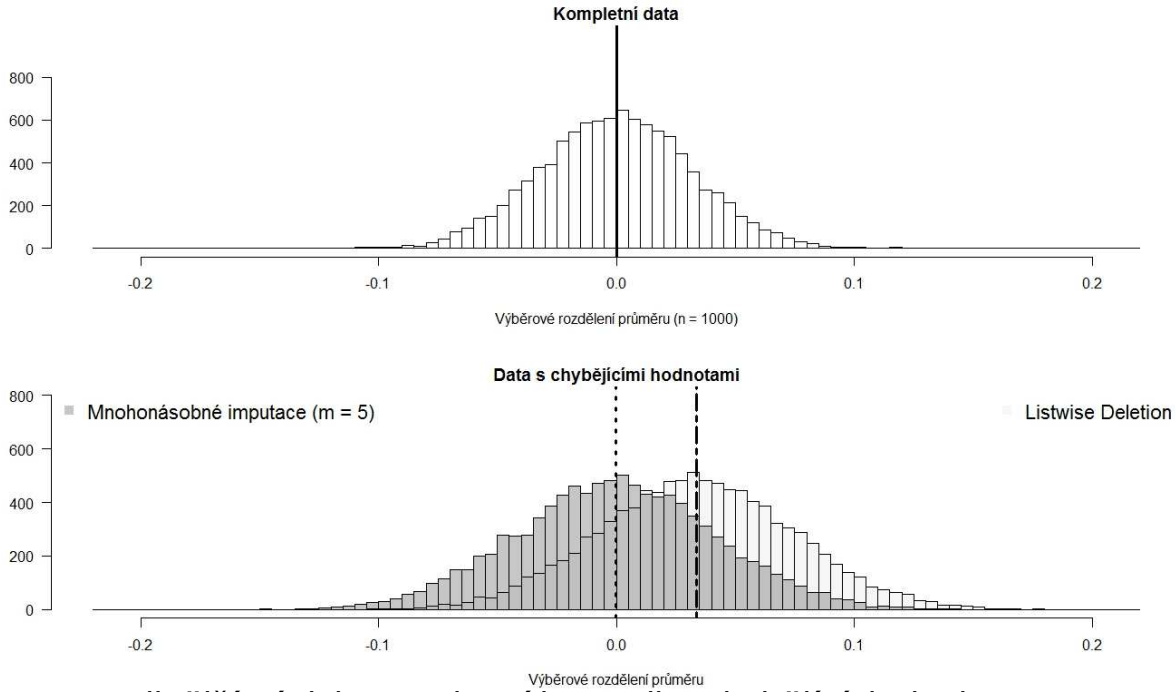
Mnohonásobné imputace

- ✓ Vedou k **nevychýleným odhadům parametrů u mechanismů MCAR a MAR**
- ✓ Pokrytí populačních parametrů intervaly spolehlivosti odpovídá zvolené hladině spolehlivosti
- ✓ **Nejlepší z dostupných metod**
- ✗ **Velmi komplexní přístup**
- ✗ Náročný proces nastavování procedury

Srovnání klíčových aspektů obou metod pomocí simulací

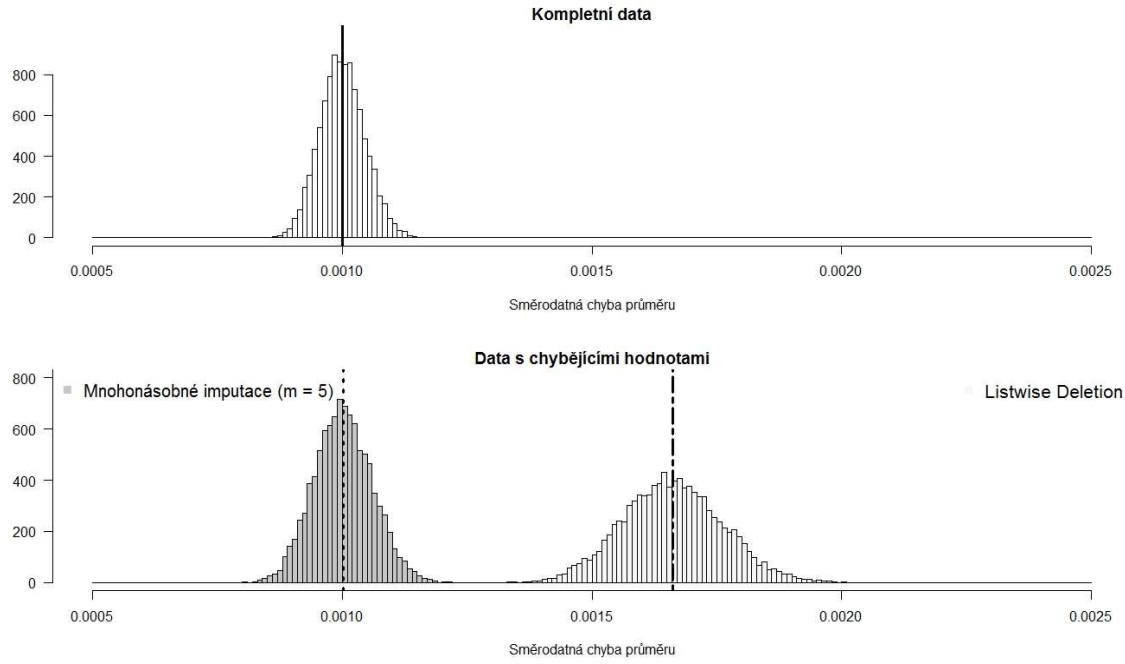
- Kompletní data byla generována ze čtyřrozměrného normálního rozdělení (**10 000 výběrů**)
- U kardinální proměnné ($\mu = 0$; $\sigma = 1$) **bylo vytvořeno 40 % chybějících hodnot** podle mechanismu MAR
- Srovnáváme výběrová rozdělení průměru a směrodatné chyby průměru získané:
  1. na **kompletních datech**
  2. s použitím **listwise deletion**
  3. s použitím **mnohonásobných imputací**

Odhadování průměru (MAR)



→ → → silnější závislost vedoucí ke vzniku chybějících dat by posunula histogram listwise deletion doprava

Směrodatné chyby průměru (MAR)



Děkuji za pozornost

ivan.petrusek@soc.cas.cz