Are we missing something with complete register data?

Statistical days, 18.5.2017 Turku

REIJO SUND

Yes, we are always missing something

- Data are only an extremely limited representation of reality / phenomenon
- How to
  - Conceptualize
  - Observe
  - Operationalize
  - Measure
  - Record systematically in symbolic form

- Data are statisticians' playground / sandbox / kingdom

Register data are secondary data

- Data are originally produced for some other purposes than the research in which it will be utilised
- There is no more possibility to tailor data collection so that it would correspond the needs of the research
- Available data may or may not be suitable for the purposes of the research
- In virtually any case require a lot of preprocessing before actual analyses

Missing data problem?

- ”Gaps” in the most common structural presentation of data
  - Some variables have ”missing values” for certain observations in a table
- Implicit assumption that data in a variable are measured in the proximity of some fixed time point
  - Only one or a few time points in cross-sectional and panel data designs
- Reality are more than a few variables in fixed time points

Phenomenon matters

- Stay constant in time
- Change systematically with time
- Change ”continuously” in time
- Change state at some time points
- …

- Event history data captures changes of states in all time points in which changes occur

Complete register data?

- Register data may be complete in the sense that similar longitudinal event-based data are available for the whole unselected population
- All (often administrative) events can be captured
  - Are all event related data recorded for each event is a different (but still certainly relevant) question
- In principle, problems of drop-out and non-response are no more major issues
- Data are typically in ”long format” as there may be a lot of time points
  - No traditional ”missing data gaps” in a table

Conclusions

- Why should we talk about missing data at all if we are actually missing most of the reality with any data?
  - Filling the gaps in a table may help in some study designs, but only on condition that the table as a structure is not missing something important
- Would it be better just to focus on the modelling of reality using available data?
  - Take into account that there may be data from any points in time
    - Feasible with event-based (register) data