Applied Data Analytics

Data analysis — Interpretation challenges

Selection problems: Introduction

Hans-Martin von Gaudecker and Aapo Stenhammar

Bob refuses to report his income

Name	Income
Alice	3000
Bob	
Charlie	5000

Q: What is mean / median income in this dataset?

Three strategies for answers

1. We don’t know *(propagate missing values)*
2. 4000 *(just ignore)*
3. Come up with a number for Bob based on external information *(impute)*

Selection: Why is data missing?

Causal question!

Goal:

- Raise awareness, provide a framework to think about it
- Constructive solutions: Later courses

Selection: Why is data missing?

1. Answer: randomly
   - No problem
   - Dropping / imputing observations tend to lead to the same result
2. Answer: for other reasons
   - Need to think hard about the selection process
   - Causal models for the selection process

Examples

- Learning from successful founders (case studies, any retrospective study)
- Polling people who spend lots of time answering polls
- Comparing health outcomes of hospitalised and non-hospitalised to learn about the effect of hospitalisation

Consequences

- Biased means, medians, variances, etc.
- Biased relationships between variables (correlations, CMF / OLS coefficients)
- Biased causal effects