Applied Data Analytics

Statistics — Basics & location

Missing data

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)*

Reasons for why data might be missing

- Refusal to answer
- Does not apply: Ask only those who are employed about labour income
- Question routing
- Privacy concerns (e.g., small cells)