Applied Data Analytics

Statistics — Basics & location

The importance of data sources

Hans-Martin von Gaudecker and Aapo Stenhammar

Why care about data sources?

Example: CPS asks some people about income, etc.

Has to be a **random sample** from the population of interest

- Ask people at the unemployment office: would not see much at the upper end.
- Ask people at the country club: would not see much at the lower end.
- More subtle than that in reality

How to get a random sample?

- If registers of the population (citizens, firms, ...) are available, draw from there
  - Some people will refuse to participate
  - But you can see who they are
- Else:
  - Random address sampling
  - Random digit dialing
  - Lots of techniques, very expensive

Examples for non-random samples

Population of interest: all people in Germany

- Asking customers on a website
- Commercial Internet surveys (with self-signup)
- Asking people on the street

Rule of thumb

If a study/fact claims that it is based on *representative data*, it probably is not based on a random sample.

Selection as a missing data problem

- If the data source is not representative, the selection process causes a missing data problem.
- More subtle than data that is missing only for some variables, because there is no way to see *in the data* that you are missing something.