Selection Bias

Expert Testimony

Six different ways of getting data

Data measured by	Within organization	Outside organization
Machine	Instrumentation, OLTP, logs, IoT	Data product, API, web scraping
Human	Survey	Survey, Statistics
Both	Labeling/annotation	Data enrichment

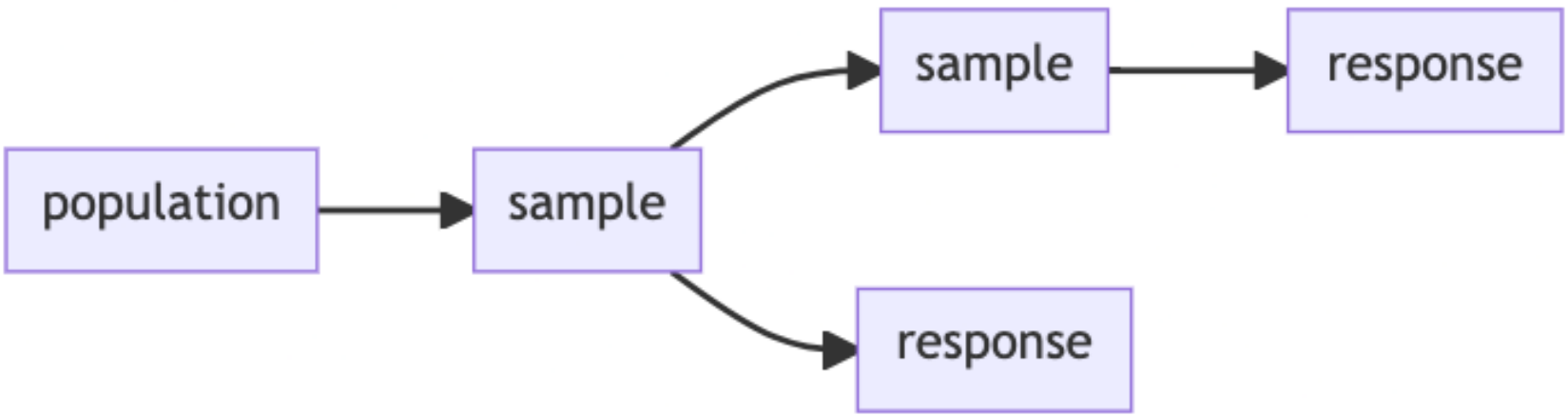
These have different implication for statistics, engineering and management.

Selection bias

If sample is not representative, may suffer from **selection bias**.

1. nonrandom selection into sample
2. incl. nonrandom attrition
3. nonrandom response rate

Getting a representative sample



What can go wrong?

1. The relevant employees are not found.
2. Sampling is is not representative.
   - ordering
   - convenience sample
3. Some (which?) employees leave the firm.
4. Some (which?) employees do not respond to survey questions.

Jargon busting

Bias

The *mean* of the data is different from what we expect. May come from selection, survival, attrition, nonresponse.

Error

Data is *noisy*, our estimate is noisy. Given proper statistical techniques, we can handle noise.