Data and Sampling Distributions

Random Sampling and Sample Bias

- In the era of big data, we still need sampling to work efficiently with a variety of data.

- Random sampling is a process in which each available member of the population being sampled has an equal chance of being chosen at each draw. The resulting sample is called simple random sample.

- Data science mostly care about the quality of the data. But statistics also cares about the representativeness of the data, in other words, how good our data in hand represent the population.

- Self-selection bias occurs when people who choose to participate in a study are not representative of the population you're trying to study.

Though it can be mitigated to a certain degree if compared to a similarly biased sample.

Bias

- Bias comes in different forms, and may be observable or invisible. When a result does suggest bias (by reference to a benchmark or actual values), it is often an indicator that a statistical or machine learning model has been misspecified, or an important variable left out.

- There are variety of methods to achieve representativeness, but most of them originates from random sampling.

- Strata is the plural form of stratum. A stratum is defined as a homogenus subgroup of a population with common characteristics. (e.g. over-weighting Hispanic and black people when sampling from a typical US state)

This is called stratified sampling.