P8130: Biostatistical Methods I
Fall 2023

Instructor: Molei Liu

COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH BIOSTATISTICS

Outline

- Introduction
- Structure and syllabus
- Goal and intention
- What are Statistics & Biostatistics
- Population, sample & types of data

Structure and syllabus

- This course is structured into modules (~ 20)
- Each module contains:
  - Required readings – textbook chapters
    1. *Rosner, B. Fundamentals of Biostatistics, 8th edition, Cengage Learning 2016*
    2. *Kutner MH, Nachtsheim CJ, Neter J, Li W, Applied Linear Statistical Models, 5th edition, McGraw-Hill International, 2013*
  - Learning checks (Quiz) – easy practice problems
  - Any extra R code documents with corresponding data
- In-person classes: Monday & Friday.
- Lectures will be recorded and uploaded to Courseworks

Structure and syllabus

- Course info can be found in the syllabus (see Canvas)

- Main expectations:

    - Class engagement
    - Complete all assignments on time; late submissions won't be accepted/graded, except for special cases (request via email before the ddl)
    - Group work is encouraged for homework, but NOT for EXAMs
    - ChatGPT: a powerful tool; use it to learn; don't copy & paste! No cheating!!

The role of homework and small quiz: helping to learn and self-evaluate; less important for grading.

Assessment and Grading

1. Learning Checks	10%
2. Homework	15%
3. Exam 1	25%
4. Exam 2	25%
5. Final Project	25%

- For final grade, pick the larger one from
  - 1 + 2 + 3 + 4 + 5
  - 3 + 4 + 5
- Will curve the grades (don’t be anxious).

Teaching Team

Liu, Molei (instructor) – ml4890

Wang, Yijin (lead TA) – yw4005

Jiao, Yixuan (A-G) – jy3269
Song, Zhengwei (H-K)– zs2539
Sun, Haochen (L-N) – hs3393
Yang, Yunxi (O-T) – yy3297
Yang, Ziyue (U-X) – zy2378
Yao, Jingyi (Y-Z) – jy3269

- Office hour: twice a week; time upon decision.
- Questions: email to the TA assigned according to the first letter of your last name.
- Welcome to contact the lead TA and instructor for any thoughts, suggestions, and special requests.

Software: R and Rstudio; Latex

- R is a FREE, open-source software used for statistical computing and graphics
  - Installing R: http://cran.r-project.org/ (Windows, Mac, Linux)
- RStudio is a user-friendly development environment for R
  - Installing RStudio: http://www.rstudio.com/
- Some online resources:
  - https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf
  - R.D., Peng Exploratory Data Analysis With R: https://leanpub.com/exdata
- If you haven’t used it, get your hand dirty ASAP!!
- Latex (or overleaf) is highly recommended (not mandatory) for HW and projects.

Goal and intention

- An introductory-level course, aimed at laying a solid foundation for subsequent learning and research.

- No advanced or cutting-edge knowledge; very few challenging math.

- (To most of you) Conversion from undergrad to grad school:

  1. Get more used to the new mode of studying;
  2. Be more professional (in your HW, project, email, etc);
  3. Be an active learner.

An active learner

- More organized and systematic. Big picture.
  - Background -> problem -> key assumption -> method -> key results -> limitation and extension.

- Be more resourceful. Online courses, notes, books, papers; Google; GPT; stack overflow, etc.
  - Never get stuck on “How to do linear regression / adjust a plot in R”!

- Put some interests on advanced or cutting-edge developments. Hasn’t to be pure method / theory.

Statistics & Biostatistics

- *Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of* ***data*** --- Wiki

- *Biostatistics is a branch of statistics that applies statistical methods to a wide range of topics in biology* --- Wiki

- Applications include, but not limited to:
  - Healthcare, Medicine, Genetics, Epidemiology, Translational Research…

- Biostatistics is not just applying statistics. It has its own methodological interests and research mode.

Statistics & Biostatistics

- A collection of observations, systematically arranged is called **DATA**
- DATA brings information as well as uncertainty
- Statistics plays important roles in
    1. Data collection
    2. Data analysis
    3. Data-driven decision-making. Hypothesis testing is a type of “decision making”

Some Definitions

- **Population:** the complete collection of units (individuals or objects) of interest in a study

  Parameter (the truth): any descriptive measure based on a population (a single-value parameter, a relationship, or a complicated function)

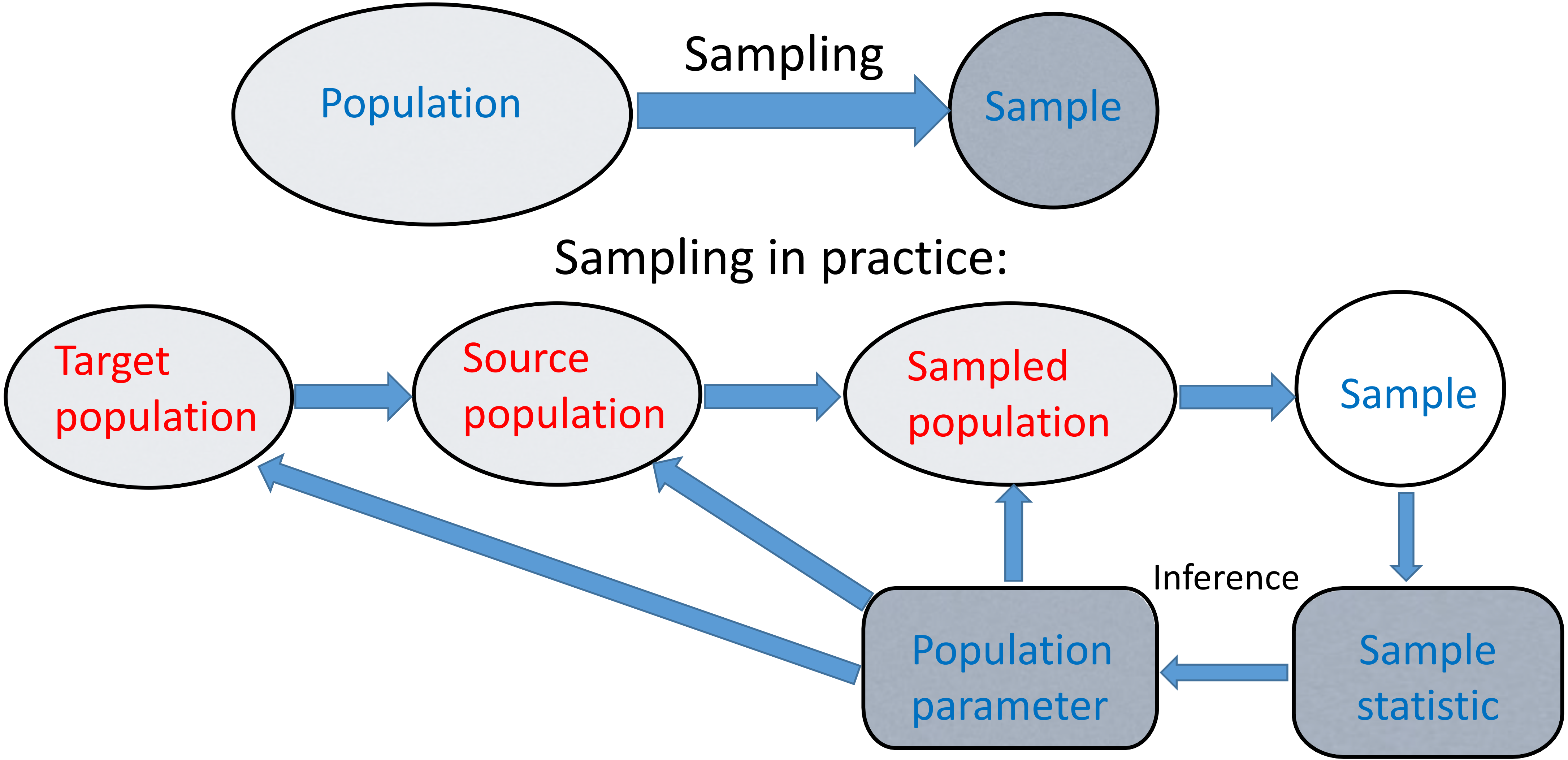
- **Sample:** a smaller subset of the population of interest

  Statistic (approximation of the truth, prone to error): any descriptive measure (of the parameter) based on a sample (synonym: Estimator)

- The **true Parameter** is invariant to the **Sample.**

- Variable: a characteristic of each element of a population or sample

Sampling Scheme



Sampling

Statistical inference allows generalization from the sample only to the population from which the sample came from.

To infer the target population, we have to make sure that the population from which the sample came from is similar to the target population.

Example:

We want to learn about all university students in NYC
We randomly sample 100 students from Columbia
We can only generalize to all Columbia students

In order to generalize to all NYC students, we have to first show that Columbia students are ‘similar’ to students from other universities in NYC.

Examples: Population-Sample

A researcher wanted to determine the prevalence of Tuberculosis in New York. She randomly selected 10000 subjects from the New York phone book and questioned them about their TB status. She found out that only 3 had TB. Describe the following

Target population:

Source population:

Sampled population:

Sample:

Parameter:

Statistic:

Examples: Population-Sample

A researcher wanted to determine the prevalence of Tuberculosis in New York. She randomly selected 10000 subjects from the New York phone book and questioned them about their TB status. She found out that only 3 had TB. Describe the following

Target population: People living in NY

Source population: All individuals in the NY phone book

Sampled population: 10,000 randomly selected individuals from the list of the individuals in the NY phone book

Sample: Individuals who agreed to be included in the study (ideally, 10,000)

Parameter: Proportion of people in NY who have TB

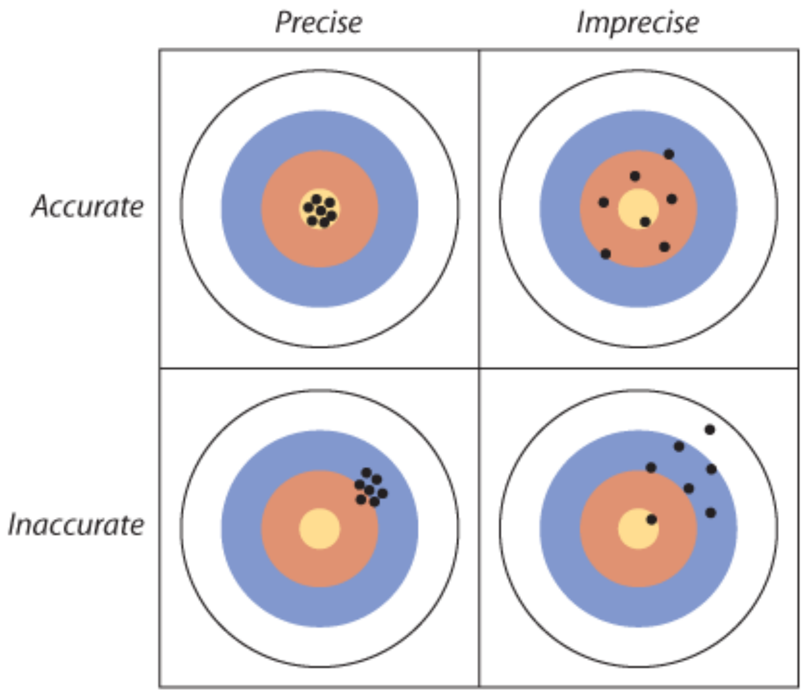
Statistic: 3 out of 9,700 (the size of sample) = 0.031%

Sampling Populations

- Reliable measures of population depend critically on how we sample populations

- Properties of good samples:
  - **Precision:** Low sampling error
  - **Accuracy:** unbiased estimates

- Bias v.s. Variance
  - **Bias**: tendency of being inaccurate
  - **Variance**: uncertainty due to randomness
  - Which is more problematic?
  - *Bias-variance tradeoff (statistical learning).



Whitlock et al, 2015

Types of Data

- Qualitative data: measurements expressed not in terms of numbers but in “types” or “categories”.

- Qualitative variables can be subdivided into:

  Ordinal variables: ordered series (e.g., preference, disease severity)

  Nominal variables: no inherent order or ranking (e.g., blood type)

  Binary variables: only two options (e.g., pass/fail, yes/no)

Types of Data

Quantitative data: measurements expressed in terms of numbers:

e.g., weight, blood pressure, survival time, etc.

Quantitative variables can also be subdivided into:

Discrete variables: usually there are gaps between the values

e.g., # of pregnancies

Continuous variables: have a set of all possible values within an interval. e.g., body mass index (BMI)

All above-introduced are structured data. Unstructured data: clinical notes, images, video, etc.

Sources of Data

- Published Source: government, business, sports statistics are collected and presented in press, online, etc.

- Experimental Study: researchers deliberately influence events and investigate the effects of an intervention

- Survey: researchers select sample of individuals and record their responses to questions

- Observational Study: researchers collect information on the attributes or measurements of interest, without influencing the events

Study Design

- Experimental

- Observational

Experimental Studies

- Randomized Clinical Trial (RCT)
  - Interventions are allocated *at random (experimental v.s. control)* and subjects are followed prospectively until an outcome is observed

- Randomization
  - Ensures that the two (or more) groups are comparable with respect to all nuisance variables (e.g., confounders)
  - In another word: we want the treatment to be independent with all potential factors influencing the outcomes.
  - *Example: testing a new diet for weight loss - want the distribution of body weight to be similar at baseline between the groups*

Observational Studies

Observational studies are to be contrasted with experiments.

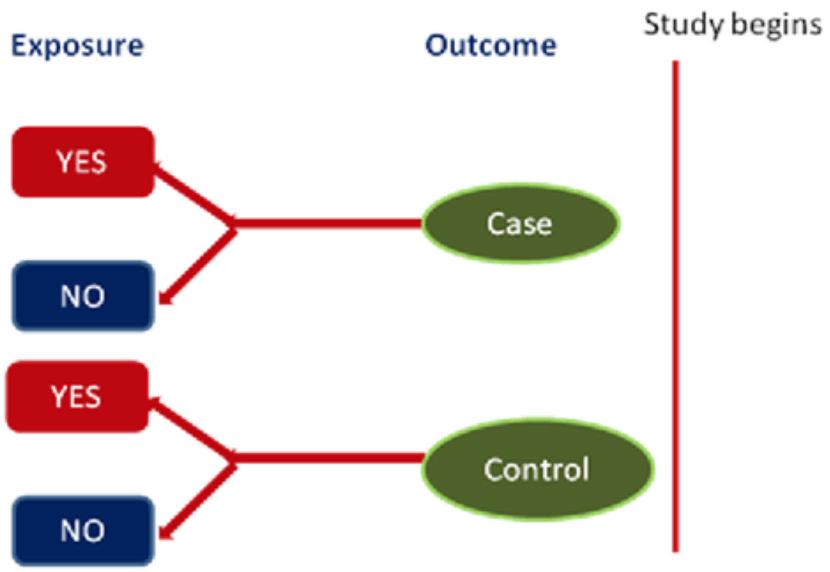
- No intervention
- Data collected on an already existing system (practical, less expensive, feasible, ethics)

Types of observational studies:

- Case study: descriptive characteristics of a *single* subject
- Case-control study
- Cross-sectional study
- Cohort study

Case-Control Study

Select subjects with disease, find matching (on potential confounders) controls and test the association between exposure and disease/outcome of interest.



PROs:

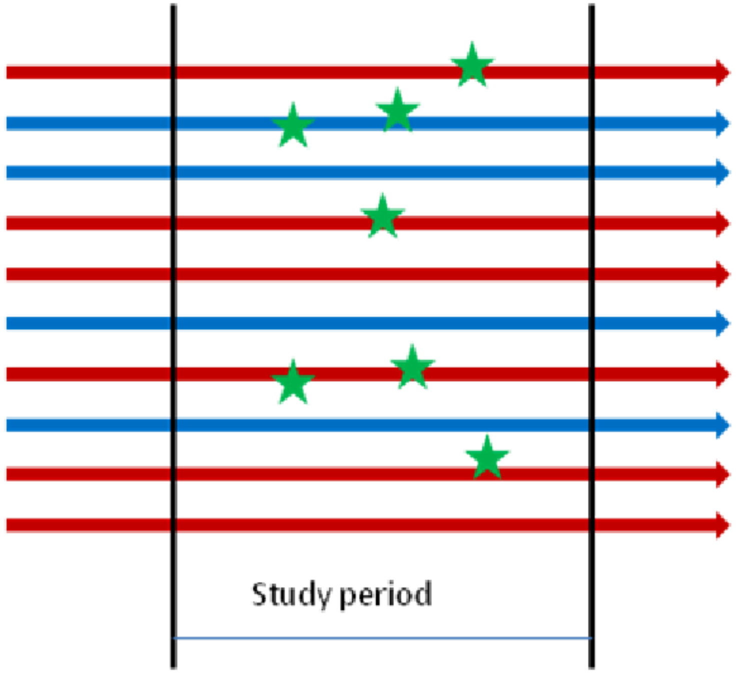
- Useful for rare diseases and long-latency
- Can explore several risk factors simultaneously

CONs:

- Difficult to select controls
- Prone to bias
- Statistical methods can get complicated (No longer completely random sampling of the target population)

Cross-Sectional Studies

Collect data from a group of subjects at one specific point in time.



PROs:

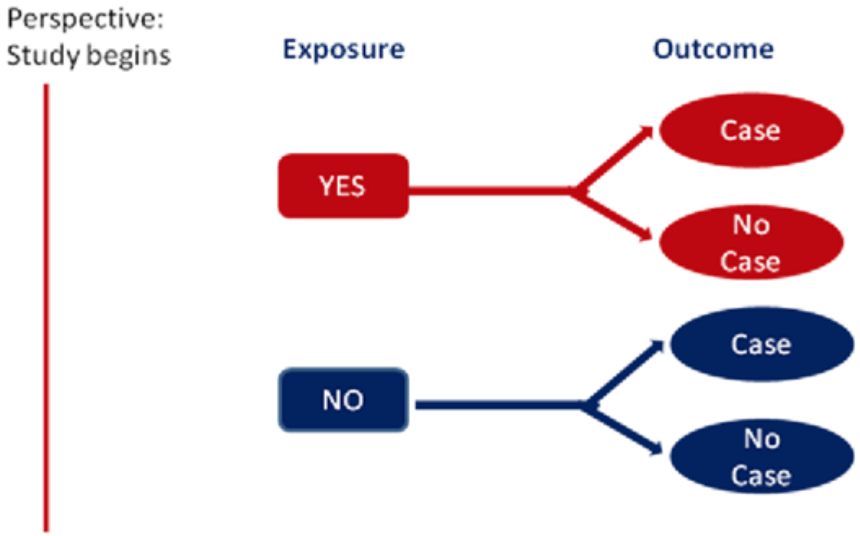
- Based on the original (target) population
- Relatively short time

CONs:

- Difficult to separate the cause-effect because of the time singularity
- Biased to determine cases with longer disease

Cohort Studies

Prospective cohort: select subjects based on exposure (or some characteristics) and follow through time to observe the development of the disease.



Retrospective: use existing data; cheaper but more restrictive in information collection.

PROs:

- Useful for rare exposures
- Cause-effect easier to establish because of temporality
- Direct measurement of incidence of disease

CONs:

- Lost to follow-up (death, drop-outs)
- Can be expensive and long