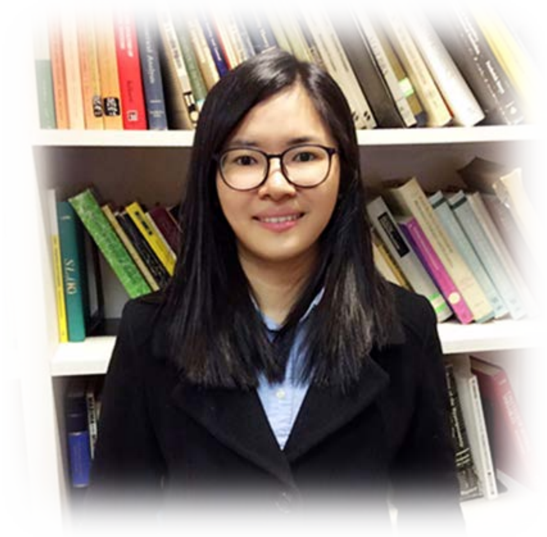
P8130: Biostatistical Methods I
Fall 2017

Instructor: Cody Chiuzan, PhD
Assistant Professor of Biostatistics at CUMC

Department of Biostatistics
Mailman School of Public Health (MSPH)

Teaching Assistants (TAs): Your New Best Friends

SAS/R Code: Examples and Practice



Zilan Chai, PhD Student

Yutao Liu, PhD Student

Theory and Applications

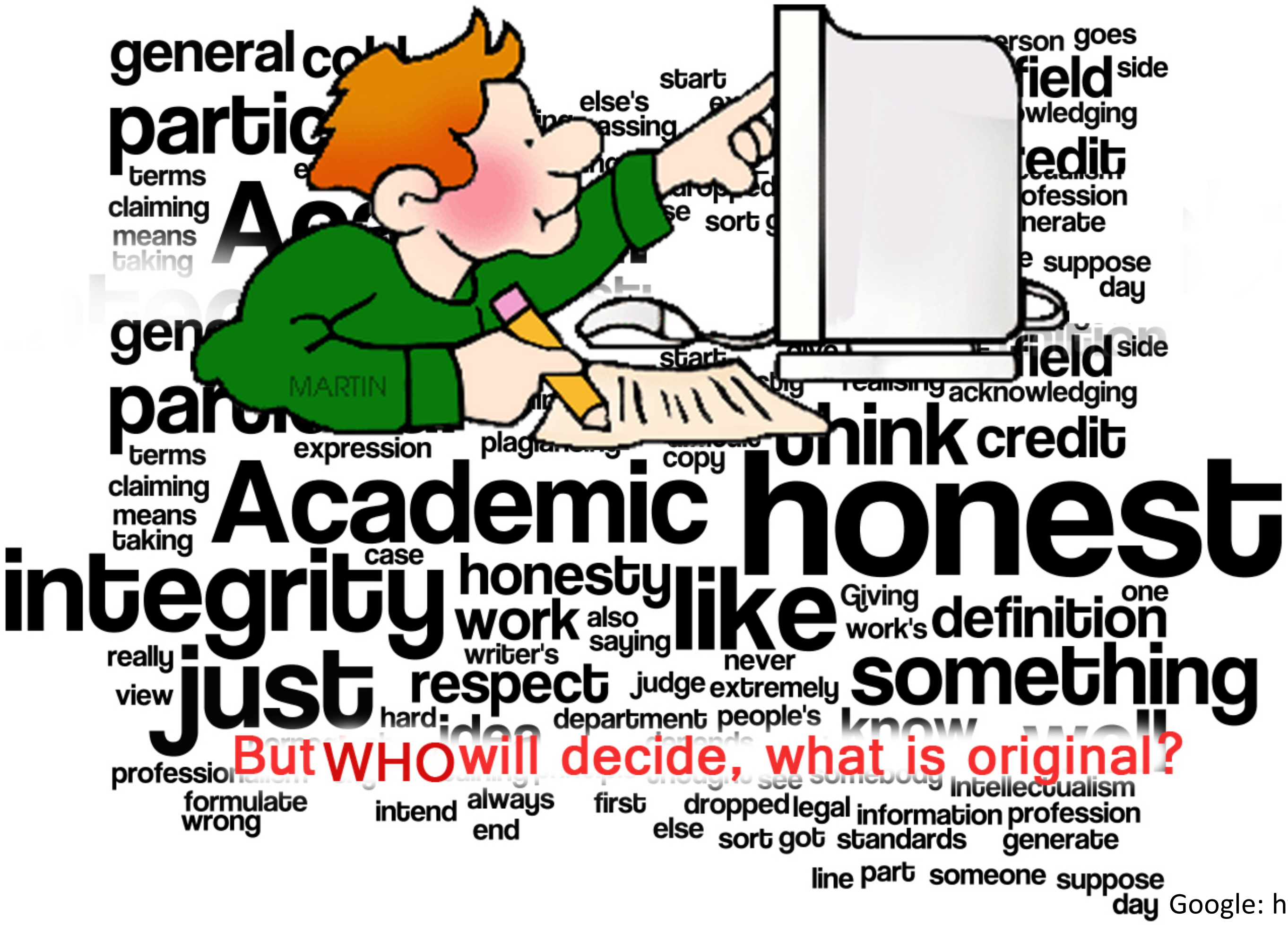
Rebecca Deek, MS Student

Shuang Wu (left), MS Student

Biostatistical Methods I (BMI)

- Course info can be found in the syllabus (see CourseWorks – new Canvas)
- Course website: XXXXXX

- Main expectations:
  - Class participation
  - Punctuality (+/- 5 min)
  - Complete all assignments on time (late submissions will not be graded)
  - Group work is encouraged for homework, but NOT for EXAMs
  - Cheating will not be tolerated!!



But WHO will decide, what is original?

Google: http://libguides.lau.edu.lb

www.turnitin.com

Internet-based plagiarism-prevention service

Software Info

- Software: R and SAS
- Make sure to download the latest R version (R 3.4.1) and also R Studio (an add-on to R)
- SAS 9.4
  - Free: MSPH computer lab - Room 656 (check availability), or other locations (e.g., Hammer Building)

    http://cumc.columbia.edu/it/students/computing-areas.html
  - Purchase: CUMC IT Service

    http://cumc.columbia.edu/it/getting_help/sas.html

Lecture 1

1. What is Biostatistics?

2. Types of Data

3. Study Design

What is Biostatistics?

The branch of applied statistics targeted towards biological and life (health) sciences.

Applied Statistics: application of statistical methods to solve real-life inspired problems involving generated data, and also develop new methodologies motivated by real problems.

Applications and contributions include, but are not limited to:

Health, Medicine, Genetics, Epidemiology, etc.

Medicine -> Biostatistics

Medicine is an empirical science, based on observations.

A collection of observations, systematically arranged is called **DATA.**

Two main components of Biostatistics:

1. Descriptive Statistics: methods of organizing and summarizing data

2. Statistical Inference: methods of making generalizations about a population based on a sample

    1. Estimation (confidence intervals)
    2. Hypothesis Testing

Some Definitions

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

Parameter: any descriptive measure based on a population

Sample: a smaller subset of the population of interest

Statistic: any descriptive measure based on a sample

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

Examples: Population-Sample

An industry sponsored clinical trial is studying a new immunotherapy for the treatment of advanced breast cancer. At the end of the study 35 out of 100 subjects participating in the trial had an objective response (reduction in tumor burden).

Variable of interest:

Population/parameter:

Sample/statistic:

What do/can we conclude?

Types of Data

Qualitative data: measurements expressed not in terms of numbers

Qualitative variables can be subdivided into:

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

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

Binary variables: only two options (e.g., pass/fail this course ☺)

Usually qualitative data are recoded as numerical values.

Types of Data

Quantitative data: measurements expressed in terms of numbers (can perform mathematical operations on data)

E.g., height, blood pressure, survival time

Quantitative variables can also be subdivided into:

Discrete variables: usually there are gaps between the values (# pregnancies)

Continuous variables: have a set of all possible values within an interval (BMI)

Sources of Data

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

Designed 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

Experimental Studies

Randomized Clinical Trial (RCT): treatments/interventions are allocated *at random* and subjects are followed prospectively until an outcome is observed

Randomization: ensures that the two (or more) groups (experimental vs control/placebo) will be comparable with respect to all nuisance variables (e.g., confounders)

E.g., testing a new diet for weight loss - want the distribution of body weight to be similar at baseline between the groups

Blinding: single and double-blind guards against placebo and experimenter effects, respectively.

E.g., researcher might be biased and measure responses differently

Observational Studies

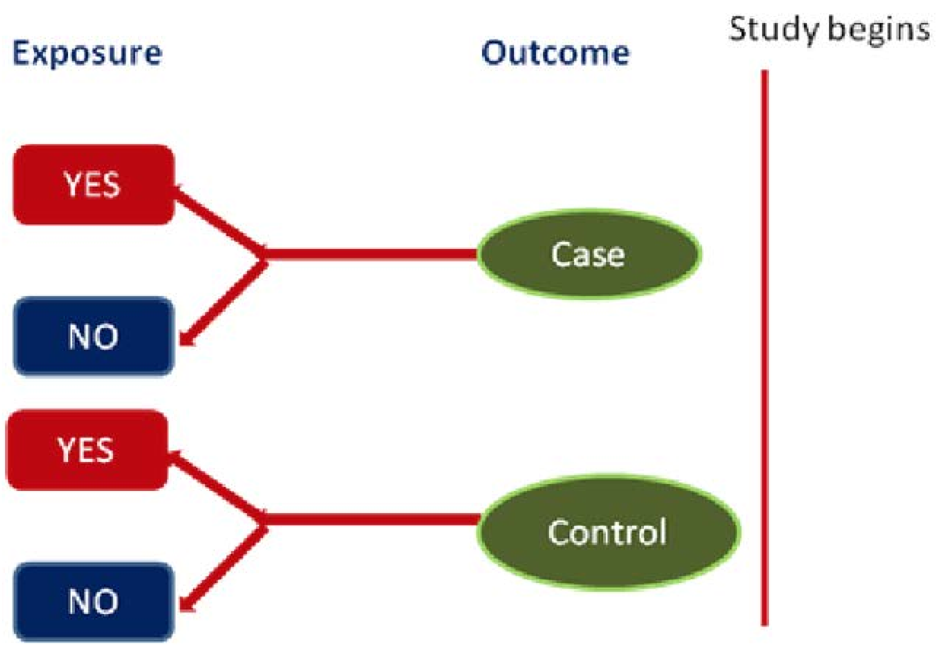
Observational studies are to be contrasted with experiments.

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

Types of observational studies:

- Case studies/case series: descriptive characteristics of a *single* subject; no comparison group
- Case-control study
- Cross-sectional study
- Cohort studies

Case-Control Studies



Backward direction: from outcome to exposure
Backward timing: study begins after outcome

Select subjects with disease, find matching controls and establish the association between exposure and disease/outcome of interest.

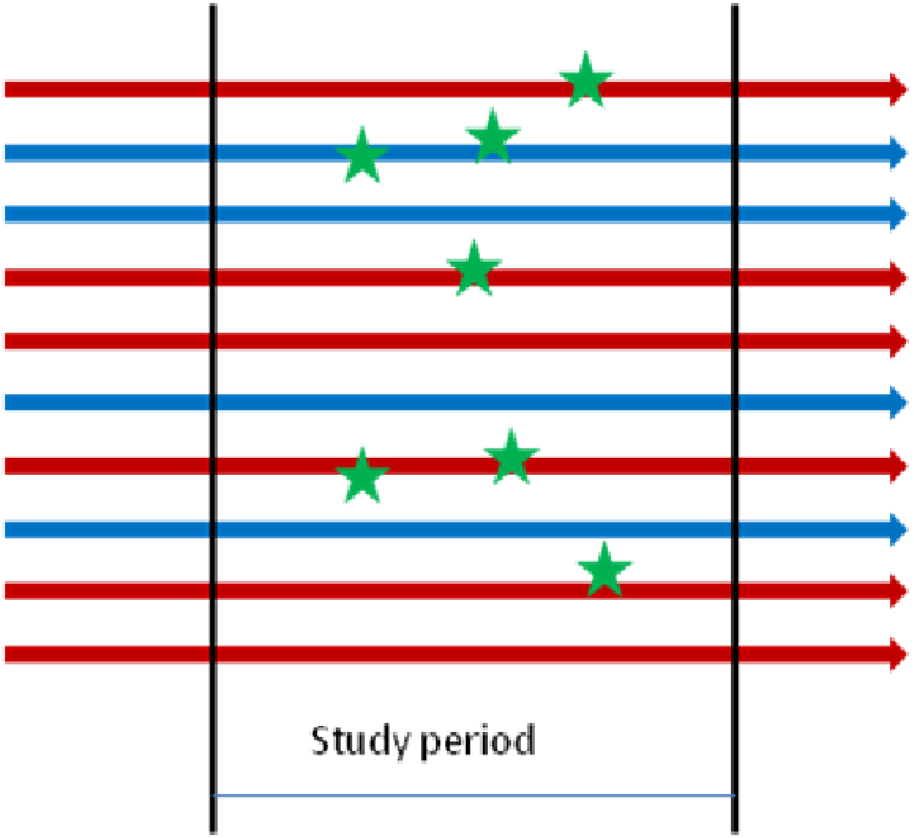
PROs:

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

CONs:

- Prone to bias
- Difficult to select controls
- Statistical methods can get complicated

Cross-Sectional Studies



Collect data from a group of subjects at one point in time. Also called prevalence studies.

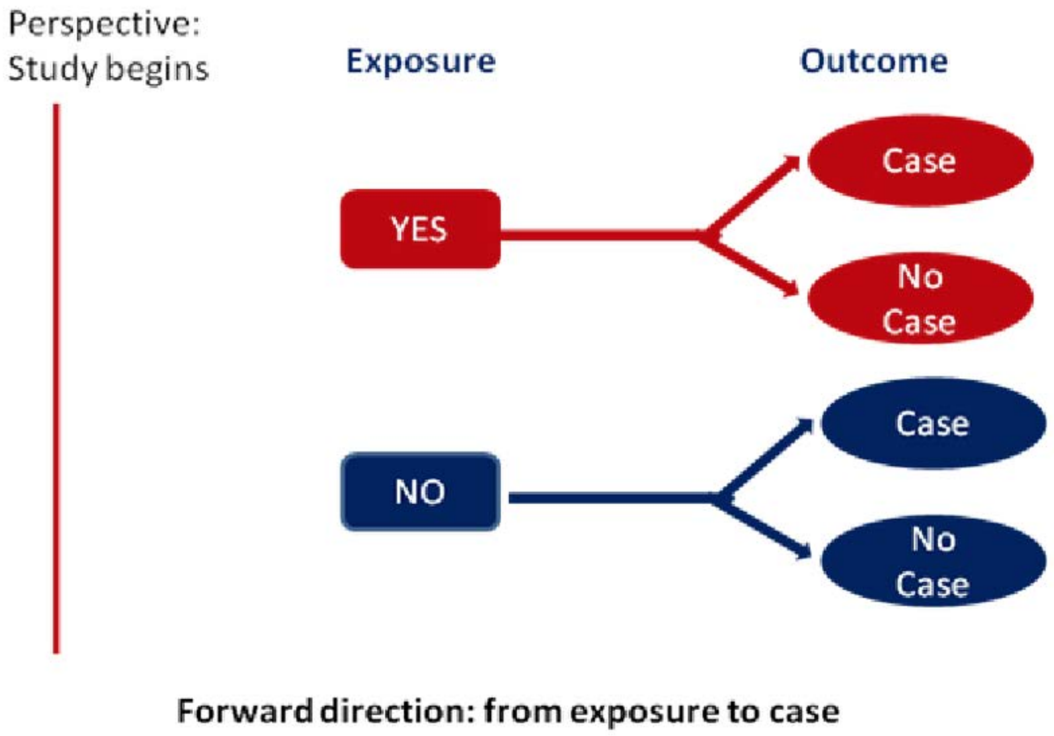
PROs:

- Bases on a sample of a general population
- Relatively short time

CONs:

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

Cohort Studies



<u>Prospective cohort:</u> select subjects based on exposure and follow through time in order to observe the development of the disease

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)
- Expensive and long

What is the role of a statistician?

We are involved in everything:

- Study design
- Power/sample size calculation
- Create the data analysis plan
- Data management
- Study implementation
- Data analysis
- Manuscript Writing

Statisticians tell or fail to tell the truth, the whole truth, and nothing but the truth…