Advanced Methods in Biostatistics

STAT 438

Winter 2022 (1221)[1]

Cameron Roopnarine[2] Yeying Zhu[3]

6th January 2022

[1]Online Course until January 27$^{\text{th}}$, 2022
[2]LaTeXer
[3]Instructor

Contents

Chapter 1

Introduction

WEEK 1
5th to 7th January

About this Course

Three topics covered in this course:

- Causal Inference.
- Missing Data.
- Measurement Error.

Basics in Biostatistics

Review:

- Experimental Studies vs. Observational Studies.
- Statistics of Interest.
- Using Regression Models.
- Association vs. Causation.

Research Questions

Questions to ask when studying a disease:

- Which factors are associated with a given disease? These so-called risk factors are sometimes referred to as predictors, explanatory variables, covariates, independent variables, or exposure variables, etc.
- Which factors are associated with the duration of a given disease?
- Correlation (Association) does not imply causation.
- Ultimately, we want to ask: which factors cause the disease, or which factors determine the duration of the disease?

Types of Studies

- Experimental studies.
- Observational studies.

1.1 Experimental Studies

- In an experimental study, the investigator can manipulate the main (risk) factor of interest, while controlling for other factors.
- In a randomized experimental study, such as a clinical trial, eligible people are randomly assigned to one of two or more groups. One group receives the treatment (such as a new drug) while the control group receives nothing or an inactive placebo.
- Due to randomization, the investigator can control for both known and unknown factors, while investigating, typically, a treatment comparison.

Randomization and Causal Inference:

- Randomization is the perfect/golden design for causal inference.
- Random assignment of treatment (exposure) ensures balance across study arms with respect to observed and unobserved risk factors.
- Direct comparisons between treatment groups can be made.
- Any difference can be attributed to the causal effect of treatment.
- Randomization is not always feasible due to ethical/economic reasons.
- Even the treatment is randomized, the participant may not comply with the assigned treatment: compliance issue.

1.2 Observational Studies

- These studies are typically based on sampling populations with subsequent measurement of various factors of interest. In this setting, we cannot even take advantage of a naturally occurring experiment that changed risk factor status conveniently.
- It is sometimes useful to use these studies to look at the natural history of a disease, but any attempt to identify causality between a risk factor and outcome must be done with great caution.
- There is no experimental setting, as study participants typically self-reflect their exposure categories. Nevertheless, in large part due to ethics, such studies are most often to what we have access in Biostatistics.

  Examples of Observational Studies

  1. - **Risk factor**: cigarette smoking.
     - **Outcome**: bladder cancer.
  2. - **Risk factor**: distance of home from hazardous waste site.
     - **Outcome**: respiratory disease.

- Three most popular observational studies:

1. Cross-sectional studies.
2. Cohort studies.
3. Case-control studies.

- No control over which subjects have the exposure and which do not.
- Exposed and Unexposed groups may be quite different with respect to other subject characteristics.
- Differences in the outcome are not only due to the (risk) factor of interest, but also because of the masking effect of other covariates (confounders).

Confounding Issue

$$\begin{tikzcd} & \text{Gender } (X) \arrow[dl] \arrow[dr] & \\ \text{Smoking } (A) \arrow[rr] & & \text{Life Expectancy } (Y) \end{tikzcd}$$

Another Example of Confounding

$$\begin{tikzcd} & \text{Critical Condition } (X) \arrow[dl] \arrow[dr] & \\ \text{Heart Transplant } (A) \arrow[rr] & & \text{Death } (Y) \end{tikzcd}$$

1.2.1 Cross-sectional Studies

- Individuals are selected from the target population and their status with respect to the risk factor and the disease status is ascertained at the same time.
- The data represents a snapshot view of the relation between the risk factor and the event occurrence.
- Surveys are often cross-section in nature where associations are of interest and less priority is given to establishing causation.
- Advantage: cross-sectional studies are typically short.
- Disadvantage: a serious problem with such cross-sectional studies is the inability to determine whether the disease outcome or the risk factor occurred first, again this makes causal inferences more problematic or almost impossible.

1.2.2 Cohort Studies

- Cohort studies typically include obtaining two groups from a pre-determined # of individuals, one possessing and the other not possessing a risk factor of interest. Subsequent counts of cases (and non-cases) of a disease of interest are then recorded.
- Much more often than not, cohort studies are prospective, but there are retrospective (or historical) cohort studies as well.

Table representing simple cohort study with sampling based on risk-factor status:

	Disease		
Risk Factor	Present (D)	Absent (D^c)	Total
Present (E)	a	b	n_1
Absent (E^c)	c	d	n_2

- $a \sim \text{BIN}\big(n_1, \mathbb{P}(D \mid E)\big)$.
- $c \sim \text{BIN}\big(n_2, \mathbb{P}(D \mid E^c)\big)$.

1.2.3 Case-control Studies

- In case-control studies, the direction of sampling differs from that of cohort studies. Specifically, the investigator selects a pre-determined # of disease cases and non-cases (i.e., controls), then looks retrospectively to see the # of individuals with and without the risk factor in each group.
- Case-control studies are retrospective studies.

Table representing simple case-control study with sampling based on disease status:

Risk Factor	*Disease* Present	Absent
Present	a	b
Absent	c	d
Total	n_1	n_2

- $a \sim \text{BIN}\big(n_1, \mathbb{P}(E \mid D)\big)$.
- $b \sim \text{BIN}\big(n_2, \mathbb{P}(E \mid D^c)\big)$.