III Causal Inference

Ishan Nath, Lent 2024

Based on Lectures by Dr. Jieru Shi

January 23, 2025

Contents

-1.1 Overview

Why causality?

- Investigators are often interested in relationships between variables.
- The major part of classic statistics is about association rather than causation.
- Association does not imply causation.

-1.2 Commonly-Used Measures of Association

1. Correlation and regression. The Pearson correlation coefficient between random variables Z and Y is
$$\rho_{ZY} = \frac{\mathrm{Cov}(Z,Y)}{\sqrt{\mathrm{Var}(Z)\,\mathrm{Var}(Y)}}.$$
This measures the linear dependence of Z and Y. Note we may get different β with different inputs:
$$\begin{aligned} Y &= \alpha + \beta Z + \epsilon, \\ Y &= \alpha' + \beta' Z + \gamma X + \varepsilon'. \end{aligned}$$
2. Contingency tables. The risk difference of a contingency table is
$$\mathrm{RD} = \mathbb{P}(Y=1|Z=1) - \mathbb{P}(Y=1|Z=0) = \frac{\rho_{11}}{\rho_{11}+\rho_{10}} - \frac{\rho_{01}}{\rho_{01}+\rho_{00}}.$$
We also have risks ratio, and odds ratio.
3. Simpson's paradox.

-1.3 Basic Terminology

Consider a study with n experimental units indexed by $i = 1, 2, \ldots, n$.

1. The intervention is (Z_i). This is what could be changed.
2. Potential outcomes and the counterfactual are: $Y_i(1)$ and $Y_i(0)$.
3. Observed outcomes: $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$.
4. Covariates (X_i): variables we observe other than Z and Y.

We have two assumptions together by the *stable unit treatment value assumption*:

SUTVA:

1. No interference: Unit i's potential outcomes do not depend on other units' treatment.
2. Consistency: If $Z_i = z$, then $Y_i(z) = Y_i(Z_i) = Y_i$.

For a binary treatment, under SUTVA, the *individual causal effect* or *average causal effect* can be defined as:

$$\tau_i = Y_i(1) - Y_i(0),$$
$$\tau = \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0)).$$

The fundamental problem of causal inference is that we can only observe one of the potential outcomes.

Causal data is a missing data problem.

Index