Lecture 1 - Introduction to Causality

DS4005 Causal Inference

Department of Statistics, University of Colombo

21 May, 2024

Identify whether the following questions are questions of **causality** or **association**

- A medical researcher wishes to find out whether a new drug is effective against a disease. causality
- A researcher wants to study the relationship between ice cream sales and drowning incidents in the US. association
- An economist is interested in uncovering the effects of a job-training program on an individual's employment prospects. causality
- A sociologist is concerned about the effects of divorce on children's subsequent education. causality
- A student wants to investigate whether there is a relationship between social media usage and feelings of loneliness among adolescents. association
- A person claims that his headache went away sooner because he took an asprin. causality

Discuss with your group members and upload your solutions via this link

Measures of association

Correlation and regression

- The Pearson correlation coefficient between two random variables A and Y is:

$$\rho_{AY} = \frac{Cov(A, Y)}{\sqrt{Var(A)Var(Y)}}$$

- The linear regression of Y on A is:

$$Y = \beta_0 + \beta_1 A + \epsilon$$

where $E(\epsilon) = 0$ and $E(\epsilon A) = 0$. Then,

$$\beta_1 = \frac{Cov(A, Y)}{Var(A)}$$

Measures of association

Contingency tables

- For two binary random variables A and Y, the contingency table is,

	$Y=0$	$Y=1$
$A=0$	p_{00}	p_{01}
$A=1$	p_{10}	p_{11}

- Let A be the exposure and Y be the outcome. In epidemiology the following measures are used to quantify the association between A and Y.
  - **Risk Difference (RD)**

$$RD = P(Y=1 \mid A=1) - P(Y=1 \mid A=0) = \frac{p_{11}}{p_{11}+p_{10}} - \frac{p_{01}}{p_{01}+p_{00}}$$

  - **Risk Ratio (RR)**

$$RR = \frac{P(Y=1 \mid A=1)}{P(Y=1 \mid A=0)} = \frac{p_{11}}{p_{11}+p_{10}} \Big/ \frac{p_{01}}{p_{01}+p_{00}}$$

  - **Odds Ratio (OR)**

$$OR = \frac{P(Y=1 \mid A=1)/P(Y=0 \mid A=1)}{P(Y=1 \mid A=0)/P(Y=0 \mid A=0)} = \frac{p_{11}p_{00}}{p_{10}p_{01}}$$

Simpson's Paradox

Definition from Stanford Encyclopedia of Philosophy

Simpson's Paradox is a statistical phenomenon where an association between two variables in a population emerges, disappears or reverses when the population is divided into subpopulations. For instance, two variables may be positively associated in a population, but be independent or even negatively associated in all subpopulations.

What is Causal Inference?

Causal inference is the study of drawing inferences about causal relationships using observed data.

Terminology and Notations

- **Unit** - A person or any other object on which the treatment is applied.
- **Exposure/Treatment/Intervention (A)** - We often consider two levels.
  - 1 - "treatment"
  - 0 - "control"
- **Outcome (Y)** - Outcome of interest.
- **Covariates/Confounders (X)** - Other measured variables in the study.

Example on taking asprin to relieve headache

- The exposure has two levels. The "treatment" ($A = 1$) level is taking asprin and the "control" ($A = 0$) level is not taking asprin.
- The outcome (Y) is the time taken to relieve the headache.
- Let's assume there are no other variables measured about the person. Hence, no other covariates X.
- We only know what happened after taking asprin. What will happen if the person did not take asprin?
- The outcome linked to each level of the exposure is called a **potential outcome**.
  - $Y(1)$ - the outcome that would have been observed if the person took asprin
  - $Y(0)$ - the outcome that would have been observed if the person did not take asprin

Next... **Potential Outcomes Framework**

Thank you