STAT931: Causal Inference and Epidemiological Studies
Lecture Notes

Yiliu Cao

Sept 13, 2024

Contents

1 Introduction

1.1 Quantities of Interest

In Epidemiological Studies, we want to measure the occurence. More specifically, we want to measure the **prevalence** and **incidence**.

Prevalence

There are two measures of prevalence

1. **Point Prevalence**: The proportion of at-risk population affected at a specific time **point**. Formally:
2. **Period Prevalence**: The proportion of at-risk population affected at a specific time **period**.

Incidence

There are also two measures of incidence.

1. **Incidence Proportion**: The proportion of a defined at-risk population who has affected within a specific time **point**.
2. **Incidence Rate**:

1.2 Association

Association is **NOT** causation, it is the comparisions between groups.

Definition 1.2.1. Relative Risk of an outcome Y with a binary risk factor A is:

$$RR = \frac{P(Y = 1 \mid A = 1)}{P(Y = 1 \mid A = 0)}$$

where $0 < RR < \infty$

Note that:

1. $P(A = 1) + P(A = 0) = 1$
2. $RR = 1 \implies Y = 1 \perp A = 1 \implies$ No association between Y and A
3. If $RR > 1$, there is greater risk of $Y = 1$ when $A = 1$ vs $A = 0, vice - versa$
4. $\frac{P(Y=1|A=1)}{P(Y=1|A=0)} \neq \frac{P(A=1|Y=1)}{P(A=1|Y=0)}$

Definition 1.2.2. Odds Ratio

- The odds of a disease Y among the exposed $A = 1$

$$= \frac{P(Y = 1 \mid A = 1}{P(Y = 0 \mid A = 1)}\Bigg) = \frac{P(Y = 1 \mid A = 1)}{1 - P(Y = 1 \mid A = 1)}$$

- The odds of disease Y among the unexposed $A=0$

$$=\frac{P(Y=1\mid A=0)}{P(Y=0\mid A=0)}\frac{P(Y=1\mid A=0)}{1-P(Y=1\mid A=0)}$$

- The odds ratio for measuring the association of disease with the exposed vs. unexposed groups is:

$$\begin{aligned}OR &= \frac{P(Y=1\mid A=1)/P(Y=0\mid A=1)}{P(Y=1\mid A=0)/P(Y=0\mid A=0)}\\ &= \frac{P(Y=1\mid A=1)/[1-P(Y=1\mid A=1)]}{P(Y=1\mid A=0)/[1-P(Y=1\mid A=0)]}\\ &= \underbrace{\frac{P(Y=1\mid A=1)}{P(Y=1\mid A=0)}}_{RR}\times\frac{P(Y=0\mid A=0)}{P(Y=0\mid A=1)}\end{aligned}$$

Note that:

1. $OR=1 \Leftrightarrow log(OR)=0$ means no association between Y and A
2. $OR\leq 1$ means greater odds of ratio when exposed
3. The OR for Y given A is equal to the OR for A given Y

$$OR=\frac{\frac{P(Y=1|A=1)}{P(Y=0|A=1)}}{\frac{P(Y=1|A=0)}{P(Y=0|A=0)}}=\frac{\frac{P(A=1|Y=1)}{P(A=0|Y=1)}}{\frac{P(A=1|Y=0)}{P(A=0|Y=0)}}$$

4. OR is a goods estimate of RR for rare disease as the second term above will approximate to 0.
5. $RR>1 \implies OR>RR$, $RR<1 \implies OR<RR$

Definition 1.2.3. Risk difference

$$RD=P(Y=1\mid A=1)-P(Y=1\mid A=0)$$

where $-1<RD<1$

This can be understood as the additional risk when exposed. In addition, a positive RD means a greater risk when exposed, vice-versa.

Definition 1.2.4. Attributable Risk(AR)
AR is the fraction of the cases of the outcome $Y=1$ that can be attributed to $A=1$

$$AR=\frac{P(Y=1)-P(Y=1\mid A=0)}{P(Y=1)}$$

Furthermore:

$$\begin{aligned}AR &= \frac{P(A=1)[P(Y=1\mid A=1)-P(Y=1\mid A=0)]}{[P(Y=1\mid A=1)P(A=1)]+[P(Y=1\mid A=0)P(A=0)]}\\ &= \frac{P(A=1)[RR-1]}{[P(A=1)RR]+P(A=0)}=\frac{P(A=1)[RR-1]}{(P(A=1)[RR-1])+1}\end{aligned}$$

We can see that the AR depends on both the association between A and Y via RR, and the prevalence of risk factor at $A=1$

Note that:

1. $AR = 0 \iff RD = 0 \iff OR = 1 \iff RR = 1$ is null, meaning that $A \perp Y$
2. $AR > 0$ means the A will rise the risk of Y; $AR < 0$ means A is protective, decrease the risk of Y.
3. AR does not imply causation.

1.3 Causation

Two variables are causally related if changing one can potentially change the level of the other. In causal inference, A represents the treatment, where $A = 1$ stands for the exposed or treated group, and $A = 0$ represents the unexposed or control group. The variable Y is the observed outcome.
In contrast, we also have the concept of the **potential outcome**, which captures the outcome under different treatment conditions.

Definition 1.3.1. Potential Outcome and Individual Causal Effect (ICE)
The potential outcome framework includes outcomes that may or may not be observed. More specifically:

- Y^0 (or $Y(0)$) is the potential outcome if assigned to the control group ($A = 0$).
- Y^1 (or $Y(1)$) is the potential outcome if assigned to the treatment group ($A = 1$).

Importantly, in reality, only one of Y^0 or Y^1 is observed for any individual; the other is counterfactual. This is also called the **Foundamental Problem**.

The **Individual Level Causal Effect (ICE)** is defined as:

$$ICE = Y^1 - Y^0$$

Since we can only observe one of Y^1 and Y^0 for any individual, this is known as the **consistency assumption**. Formally, the observed outcome Y is represented as:

$$Y = AY^1 + (1 - A)Y^0 \quad \text{or} \quad Y = Y^A$$

This implies that an individual cannot belong to both the treatment and control groups simultaneously. However, an individual can receive different treatments at different time points.

Remark 1.1. Stable Unit Treatment Value Assumption (SUTVA)
The consistency is implied by the **Stable Unit Treatment Value Assumption (SUTVA)**, which consists of two key conditions:

1. **No interference**: The outcome of one individual is not affected by the treatment status of other individuals.
2. **Well-defined treatment**: There is a single version of the treatment; no multiple versions of the treatment exist.

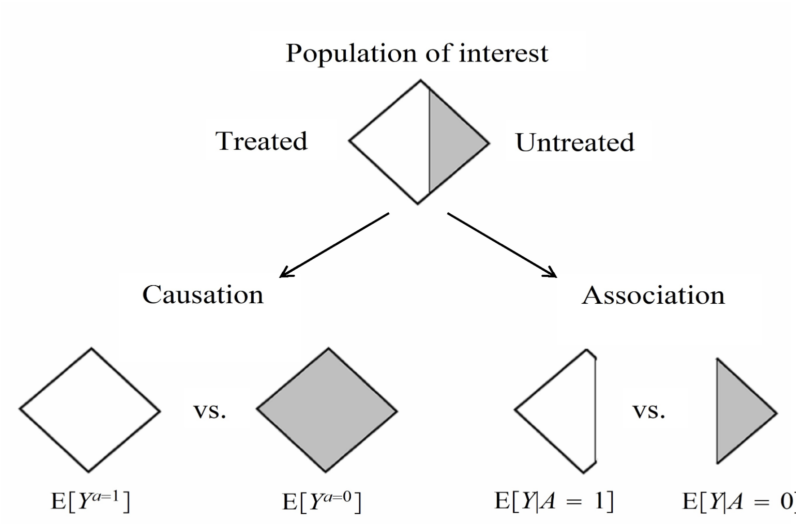

Figure 1.1: Visualization of Association and Causation

Definition 1.3.2. Average Causal Effect (ACE)
To solve the foundamental problem, instead of using ICE, we take ACE.

$$\begin{aligned} ACE &= E\left[Y^1 - Y^0\right] \\ &= E\left[Y^1\right] - E\left[Y^0\right] \\ &= 1 \times P[Y^1 = 1] + 0 \times P[Y^1 = 0] - (1 \times P[Y^0 = 1] + 0 \times P[Y^0 = 0]) \\ &= P\left[Y^1 = 1\right] - P\left[Y^0 = 1\right] \end{aligned}$$

1.4 Association vs. Causation

Remark 1.2. $E[Y^1] \neq E[Y|A = 1]$

- $E\left[Y^1\right]$ is a priori, and is the average value defined by the **entire population**.
- $E[Y \mid A = 1]$ is defined after treatment assignment, and averaged only **among** those who receive treatment $A = 1$.
- We use $E\left[Y^1\right]$ in Causation and $E[Y \mid A = 1]$ in Association.

Figure 1.1 visualizes the difference between association and causation. Moreover, like discussed above, we also have causal **Risk Ratio**, causal **Odds Ratio** and causal **Risk Difference**.

Definition 1.4.1. Measures of Causal Effect

1. **Causal Risk Ratio**

$$RR = \frac{P\left(Y^1 = 1\right)}{P\left(Y^0 = 1\right)}$$

2. **Causal Odds Ratio**

$$OR = \frac{P\left(Y^1 = 1\right) / \left[1 - P\left(Y^1 = 1\right)\right]}{P\left(Y^0 = 1\right) / \left[1 - P\left(Y^0 = 1\right)\right]}$$

3. **Causal Risk Difference**

$$RD = P\left(Y^1 = 1\right) - P\left(Y^0 = 1\right)$$

Remark 1.3. Summary of Association vs. Causation

1. (Associational) Risk Ratio	**1. Causal Risk Ratio**
$RR = \frac{P(Y=1 \mid A=1)}{P(Y=1 \mid A=0)}$	$RR = \frac{P(Y^1=1)}{P(Y^0=1)}$
2. (Associational) Odds Ratio	**2. Causal Odds Ratio**
$OR = \frac{\frac{P(Y=1 \mid A=1)}{1-P(Y=1 \mid A=1)}}{\frac{P(Y=1 \mid A=0)}{1-P(Y=1 \mid A=0)}}$	$OR = \frac{\frac{P(Y^1=1)}{1-P(Y^1=1)}}{\frac{P(Y^0=1)}{1-P(Y^0=1)}}$
3. (Associational) Risk Difference	**3. Causal Risk Difference**
$RD = P(Y=1 \mid A=1) - P(Y=1 \mid A=0)$	$RD = P(Y^1=1) - P(Y^0=1)$

The **Sharp Null/Causal Null** (for ACE) means no causal effect; equivalently $AR = 0 \iff RD = 0 \iff OR = 1 \iff RR = 1$. However, holding sharp null for ACE does **NOT** necessarily mean the sharp null for individual causal effect also holds.

2 Randomized Trials

2.1 Experimental Studies and Randomized Experiements

Experimental studies means the investigator can **manipulate** the factor of interest to observe the effect on another variable. For example, we can give drugs to one group of people and compare the outcome to another group which did not take the drug.

Randomized experiment is a type of experiment in which participants are randomly assigned to different groups, meaning that each participant has equal chance to be assigned to treatment group. The randomization can be helpful to distribute both known and unknow confoudning variables, therefore can make a stronger claim.

However, the **foundamental problem** still exist, meaning that we still can not observe both Y^0 and Y^1 for any participant.

Exchangeability

Randomized experiment ensures exchangeability, which the treatment received is **independent** of potential outcomes. i.e.

- $A \perp Y^1$ and $A \perp Y^0$
- $P\left(Y^1 \mid A=1\right) = P\left(Y^1 \mid A=0\right) = P\left(Y^1\right)$

Identification under randomization

The causal effect is equal to association under randomization

$$\begin{aligned} E\left[Y^1\right]-E\left[Y^0\right] &= E\left[Y^1|A=1\right]-E\left[Y^0|A=0\right] \text{ (by exchangeability/randomization)} \\ &= E[Y|A=1]-E[Y|A=0] \text{ (by consistency; } Y^A=Y) \end{aligned}$$

We say ACE is identified. The associational RD is an estimate of the causal RD.

2.2 Randomization methods

Bernoulli Trial

- Assign treatments based on the flips of coin.
- Yield balance in large samples
- May assign all samples to treatment (Unbalanced).

Completely randomized experiment

- Randomly assign participants into groups of fixed size, like half and half for treatment and control.
- Balance across treatment groups.

Block randomization

- Firstly divide the sample into K blocks.
- Then perform completely randomized experiment within each block.
- Ensures a balance in sample size across groups.
- A special case of CRE.

Stratified randomization

- Control and balance the influence of covariates.
- Divide the sample group to subgroups/strat based on some covariates.

3 Stratified Designs

3.1 Identification under conditional randomization

The pooling stratum-specific effects only works if there are no effect of modification. To identify the causal effects, we introduce two other methods which are **Standardization** and **Inverse Probability Weighting**. By the way, we say causal effect is identified if it can be written in terms of observed data.

3.1.1 Standardization

We would assume consistency which says $Y = Y^A$ by SUTVA and conditional exchangeability $A \perp Y^a \mid L$ for $a = 0, 1$. Then we have:

$$\begin{aligned}
E\left[Y^1\right] &= \sum_I E\left[Y^1 \mid L = I\right] P(L = I) \quad \text{(by total expectation)} \\
&= \sum_I E\left[Y^1 \mid A = 1, L = I\right] P(L = I) \quad \text{(by cond'l exchangeability/randomization)} \\
&= \sum_l E[Y \mid A = 1, L = I] P(L = I) \quad \text{(by consistency)}
\end{aligned}$$

3.1.2 Inverse Probability Weighting

We weight each outcome as the inverse