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 1.1. 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. 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 1.3. 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 1.4. 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 1.5. 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, 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.

1.3.1 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.