L1: Time in epidemiology: Questions and parameters

2022-01-13

Readings

- Samet JM. Concepts of time in clinical research. Annals of internal medicine. 2000 Jan 4;132(1):37-44.

Introduction

Today, we start with the big picture. Why do we endeavor to do epidemiologic research? What makes a study impactful? How do we go from the outputs of statistical methods to improving clinical and public health decision making?

2 goals in this course:

1. Ask good questions
2. Learn tools to answer these questions in settings with non-ideal data

Questions in epidemiology

Asking questions is the most important part of epidemiology. What makes a good question?

1. Asking the right question at the right time
2. Framing the question so that answers are meaningful

Asking the right question at the right time

Decisions in public health and clinical medicine are (typically) driven by evidence, but must be made with whatever information is available. If no one has asked the question, no information is available.

We worry a lot (though perhaps not enough) about missing data in our studies in epidemiology. But it is easy to overlook information we are missing because an important question went un-asked.

We cannot ask every question, but we must recognize our blind spots and fill them in when important.

There are many routes to asking a question, and many sources of influence

It is important to recognize that there are forces beyond individual choices that shape the questions that are asked. When we consider our blind spots, it is imperative that we keep these social forces in mind. Those with the power to determine which questions get asked also dictate the information available for decision making.

Framing questions

Appropriately framing questions ensures that questions are answerable and that answers are useful.

Types of questions (a rough taxonomy)

1. Descriptive: What happened?
2. Predictive: What will happen?
3. Causal: What would have happened if we had done X?
4. Causal-predictive: What will happen if we do X in the future?

How will results be used?

- Who will be acting on the results?
- Whom does the decision affect?
- What actions are feasible, now or in the future?
- What types of results would be compelling to decision makers?

Elements of a well-framed question

1. *Target population*: who would be affected by any decision made as a result of this study?

2. *Time period*: When is the question relevant? When would the decision affect people? To appropriately specify the time period, we must consider both the `origin` and the `timescale`.
   A. The **origin** is the timepoint at which follow-up begins. Choosing an appropriate origin is critical to ensure that results are meaningful. For studies examining effects of decisions, the origin should usually correspond with the earliest decision point. For studies reporting disease incidence, the origin should correspond with the time at which we become interested in people, either because they are the people who would be affected by a future decision or because at that time point they meet some criteria that puts them at risk of an outcome.

   i. Examples:

   - When studying the incidence of Chagas disease in childhood, the origin might be `birth`.
   - When studying the incidence of cancer among people working in a uranium mine, the origin might be `age 18`
   - When studying mortality among people with HIV, the origin might be `HIV seroconversion`

   ii. Caveat: There are usually many choices for the origin, and this choice critically affects how results are interpreted.

   B. The **timescale** depends fundamentally on the origin. If the origin is `HIV seroconversion`, the timescale is naturally `time since HIV seroconversion`. If the origin is `birth`, the timescale becomes `age`.

3. *Outcomes*: What is the relevant endpoint? When are outcomes measured with respect to the timescale?

4. *Actions*: if there are a set of actions being assessed, what is the set of candidate actions? when would they occur?

5. *Groups*: if there is a comparison being drawn between groups, the set of groups compared, including when groups are defined

Discussion: What is the difference between *Actions* and *Groups* in the above schema?

Consider the idealized cohort study to flesh out your questions

What is a cohort study?

> "... all subjects in a source population are [classified according to their exposure status and] followed over time to ascertain disease incidence" - Modern Epidemiology

What is an idealized cohort study?

- Enroll everyone in the target population
- Follow everyone from the `origin` of interest
- Measure outcomes on everyone
- If we want to compare between groups, measure group membership on everyone
- If we want to compare outcomes under various actions, we will need to either become comfortable with metaphysics (e.g., perform the action, observe the outcomes, then go back in time, do the opposite action, and observe how outcomes would have been different) or specify an idealized `target trial`.

Frame questions without considering existing data or logistics of data collection

Considering the limitations of existing data when framing a question can limit scope and lead to asking questions that are not impactful. Think about the question that needs to be asked, and then as a next step, carefully evaluate how available data (old or new) fits in to answering that question. During this step, you will almost certainly identify data gaps and assumptions you must make to answer your question with your data. This is good. Being cognizant of these gaps and assumptions prevents overstating results and informs additional data collection or sensitivity analyses.

Consider absolute measures in addition to contrasts

In this course, we will define "absolute" measures as those that can be estimated for a single sample and describe the magnitude of some health issue. Estimates of risk, prevalence, and rates are absolute measures. We often contrast these measures between exposure groups (using differences or ratios). But reporting the absolute measures themselves is helpful to inform decision making. In this class, we will often ask you to consider absolute measures in addition to their contrasts.

Always ask "compared to what"

- what is a realistic alternative to any action considered?
- are there competing events or other risks/benefits to be weighted
- methdological work should also consider alternatives (e.g., Method A might be worse than Method B, but is much better than Method C, which is used by nearly everyone in the field).

Parameters

What is a parameter?

In this course, we will use the term `parameter` to be synonymous with `estimand`. Both refer to the quantity we are trying to estimate. Example parameters might be the `risk`, `rate`, `prevalence`, or `hazard`.

Risk

> "Risk is [the] probabiliy of an event during a specified period of time" - Modern Epidemiology, page 10

Aside: what is a probability?

According to Modern Epi, a probability is

- Relative frequency of an event over time
- Tendency, or propensity, of an entity to produce an event
- Degree of certainty than an event will occur

Typically, epidemiologists define risk as the proportion of units who experienced the outcome within a defined time period.

Risk *functions*

In your training up until now, you may have seen risk defined at single time points (e.g., risk of Chagas infection by age 10). However, often, we are interested in when events occur and how risk varies over time. To allow examination of risk over time, we will broaden our definition of risk such that

> **Risk** is the probability of experiencing the event of interest at or before a given timepoint

For example, rather than estimating risk of Chagas infection by age 10, we might report risk of Chagas infection by age 1, age2, age3, ... and so on.

We will write this parameter as

$$P(T_i \leq t)$$

where T_i is the time from origin to event for person i and t is the timepoint at which we would like to evaluate risk.

Hazard

The hazard is the "instantaneous" rate of having the event at a specific time t, among participants who have not had the event prior to that time. In notation, the hazard is typically expressed as the limit of the probability of having the event in an interval of width Δt divided by Δt, as Δt goes to 0.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid t \leq T)}{\Delta t}$$

Hazards themselves are not of central interest in this course. Instead, we will focus primarily on risks because they are easier to communicate and interpret and provide more natural inputs into decision making processes. However, in settings with censoring and truncation, the hazard function can be a useful stepping stone to the risk function.

Estimators

Risk in closed cohorts

Aside: What is a closed cohort?

> A cohort where all participants are followed from the origin until the end of the risk period. That is, no participants enter the study late or become lost to follow-up.

For example, in the table below, we could compute risk at 5 years as $3/6 = 50\%$.

id	Event_before_5years
1	0
2	0
3	1
4	0
5	1
6	1

If we want to consider risk as a function over time, we need more information. Specifically, we need the times at which each participant experienced the event. Therefore, a key feature of data for "time-to-event" analyses is a dataset that includes not only *whether* or not each participant had the event, but *when* they had the event.

(Note, that the timing of the event is always important (even when reporting risk at a single timepoint), because it allows us to determine whether the individual experienced the event within the risk period. But, time is not always included in analytic datasets when the goal is to estimate risk at a single time point.)

Exercise: Consider the dataset below, which reports the time of each event. What is the risk at 1, 2, 3, 4, and 5 years after the origin? Plot estimated risk at each time point.

id	Time_of_Event
1	
2	
3	2
4	
5	1
6	4

Notation

Let's begin with risk at a single time point. Say we want to estimate risk at time 10. We can write this risk as

$$F(10) = \frac{1}{n}\sum_{i=1}^{n} I(T_i \leq 10)$$

where, again, T_i is the time from the origin to event for participant i. We can generalize this expression to multiple timepoints by substituting t for the specific time point of interest:

$$F(t) = \frac{1}{n}\sum_{i=1}^{n} I(T_i \leq t)$$

What about people who never have the event?

Some events, like death, are inevitable. We say that these events have *no competing events*. For events like death, we know that everyone in the study will eventually have the event (though the time of the event may be past the end date for the study). Other events may be prevented entirely by *competing events*. For example, death is a competing event for cancer incidence. We will discuss competing events in depth later in the semester. In settings with censoring and truncation, more complex methods are needed when competing events are present. However, in closed cohorts, we can estimate risk directly even if competing events are present using the estimator below

$$F(t, j) = \frac{1}{n}\sum_{i=1}^{n} I(T_i \leq t, J_i = j)$$

where T_i is now the time from the origin to the first of the event of interest or the competing event and J_i indicates the first type of event to occur.

For example, consider the table below. What is the risk of cancer incidence over time? Plot the risk of cancer incidence at 1, 2, 3, 4, and 5 years.

id	Time_of_Cancer	Time_of_Death
1		2
2		
3	2	3
4		
5	1	5
6	4	

Risk in open cohorts

Risk is more difficult to estimate in cohorts with late entry or loss to follow-up. In the weeks to come, we will explore estimators of risk in these settings.

For now, let's consider the setting where we would like to estimate the risk of outcome Y by 10 years after the origin, but some participants are lost to follow-up, as shown in the table below, and are therefore missing outcome data.

id	y
1	0
2	0
3	0
4	1
5	
6	

In the table above, Y is an indicator that $T_i \leq 10$. We cannot compute risk directly because we are missing outcome data for participants 5 and 6. But, we can compute *bounds* on the risk. What is the highest risk could have been, given the observed data? What is the lowest that risk could have been, given the observed data?

Line diagrams

Line diagrams are useful to visualize choices of origin and timescale and to see which participants are being compared at a given point in time.

Steps to drawing a line diagram:

1. Draw and label axes (y axis is usually study id, x axis is timescale, starting at the origin)
2. Calculate the amount of time after the origin that the first person enters the study. Place an open circle at this timepoint.
3. Draw a line from this open circle until the last available information for that participant.
4. If an event was observed, place a closed circle at that time
5. If no event was observed, place an arrow at that time
6. Repeat for remaining participants

For `R` code to produce line diagrams, please see the Resources section of the site.

Line diagram example

Say have a (small) cohort study of 5 soldiers selected at random from all people enlisting in the armed services between 2004 and 2014 and followed for mortality up to 10 years.

id	age	enlist	last_info	vital_status
1	24	2008-07-01	2022-01-12	alive
2	18	2008-07-01	2011-01-01	dead
3	21	2011-01-01	2016-06-15	dead
4	35	2011-01-01	2022-01-12	alive
5	19	2013-07-01	2022-01-12	alive

Practice drawing 3 line diagrams:

1. one with age (starting at age 18) on the x axis
2. one with calendar time (starting 1 Jan 2008) on the x axis
3. one with time since enlistment on the x axis.

For each, note the

- number of deaths occurring within the 10-year follow-up period
- number of late entries