Contact tracing can give a biased sample of COVID-19 cases

tl;dr Your sample is incomplete and your missing cases are not missing at random.

The *New York Times* today has an article entitled "Small Gatherings Spread the Virus, but Are They Causing the Surge?" It quotes several epidemiologists who answer (I'm paraphrasing) "we don't know", but the whole tone of the article suggests that we do, and that its headline question should be answered in the negative.

It presents evidence like this:

> In Colorado, only 81 active cases are attributed to social gatherings, compared with more than 4,000 from correctional centers and jails, 3,300 from colleges and universities, nearly 2,400 from assisted living facilities, and 450 from restaurants, bars, casinos and bowling alleys.

> In Louisiana, social events account for just 1.7 percent of the 3,300 cases *for which the state has clear exposure information.* [emphasis added]

But these statistics are meaningless. The reason is that it's quite hard to establish where someone got infected with COVID-19. Generally this happens as part of contact tracing, and only a small proportion of cases are currently being effectively contact traced in most states. Critically, those that are being traced are very unlikely to be a random sample of *all* cases.

But don't take it from me, take it from someone who actually works with this data:

> @brittanygrogan It seems obvious to me that people ACTUALLY working with raw local covid/contact tracing data were not consulted for this article. Because I do, and it is so. much. easier to identify & link cases from LTCFs, jails, schools, childcare, & workplaces than from private gatherings.

This is the key point. Some events or settings—like weddings, with their fixed guest list, or jails, with their inmate register—are inherently easy to trace people from.

Other settings are harder: a restaurant, for example, may have a list of bookings, but for walk-ins who pay with cash they may have no useful information with which contact tracers can work. Some settings, like a grocery store or bodega must require superhuman efforts to effectively contact-trace from. In a system with stretched resources, any individual home gathering of a few people is probably not going to get much attention.

So when you see that Louisiana has exposure information for 3300 cases (actually 3540 as of right now), you need to realize that that represents less than 2% of Louisana's estimated 221,000 cases. Such a tiny sample is highly unlikely to be representative.

It's even worse when you remember that contact tracers are not trying to obtain a random sample; they're trying to maximize the number of previously-ignorant infected people they can find for a given level of effort. It seems likely that that will result in a sample that is systematically biased towards certain kinds of low-effort / high-payoff settings.

We can use a toy model to show this effect. Like any model it makes a lot of simplifying assumptions, but hopefully you can see that the underlying mechanism is plausible and would still apply in more realistic circumstances.

A simple model to show how this can happen

Assume a community with 1000 people, and a single time period. In that time period, every person attends exactly one event, either:

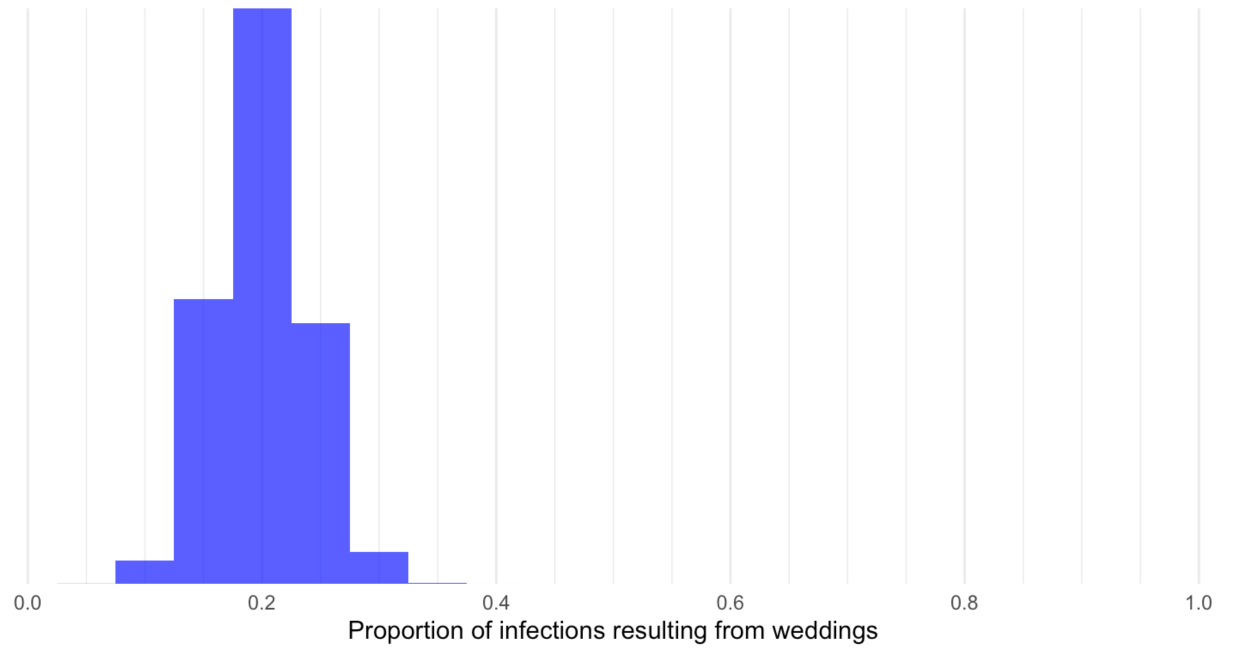
- one of 2 weddings, each of which has 100 attendees (accounting for 200 people total)
- one of 80 brunches, each of which has 10 attendees (accounting for 800 people total)

Suppose that exactly 10% of people at every event are infected, regardless of the type of event. Then this is our true "outbreak setting table," which shows that 20% of cases were a result of weddings.

Outbreak setting	Number of outbreaks	Cases	Of total cases
Wedding	2	20	20%
Brunch	80	80	80%
Total	82	100	

We can complicate that slightly by assuming that rather than exactly 10% of people at each event being infected, each person has a 10% chance of being infected (iid, if you want to be technical). That's a subtle difference; it creates some randomness, so we may not get exactly 20% of cases resulting from wedding every time.

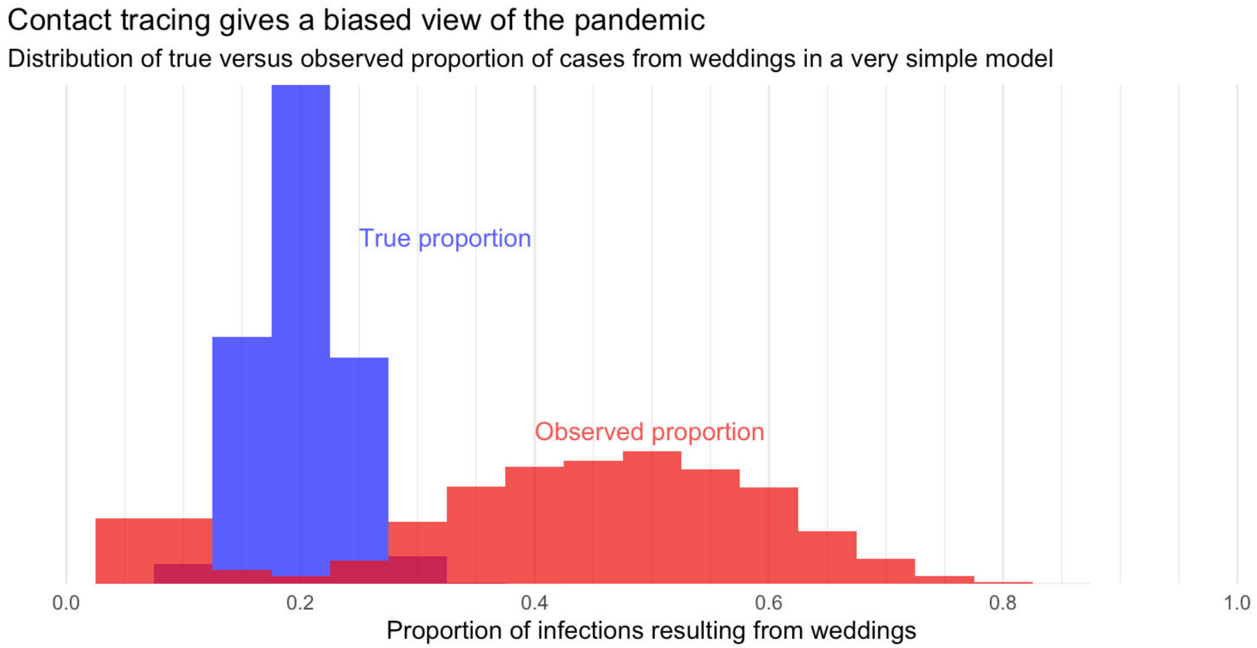
In fact, below is the distribution of that proportion, simulate over 50,000 trials. You can see it still centers on 20%, with some variation around that.



Of course the health authority can't observe this proportion or the table above directly, but has to discover it through contact tracing. Suppose that contact-tracing is imperfect, and that due to faulty recall of patients and staff shortages, an infection has only a 20% chance of being traced to a source event. Call that "primary contact tracing."

But then—and this is critical—assume there is a "secondary contact tracing" step. If two infections are independently traced to the same source event, a special effort is made to test every person who attended that event, with the result that 100% of infections associated with that event are identified. (I have no idea if public health agencies actually do this, but it seems likely.)

What will this mean for the observed proportion of cases traced back to weddings?



Under this model we're very likely to overestimate the proportion of cases that result from weddings (and equivalently, underestimate this that come from brunches). The mean observed proportion is around double the true proportion, and the modal observed value is even higher.

If you were a *Times* reporter looking at the raw data—a draw from the red histogram—you might be persuaded that weddings were, in aggregate, a more important source of transmission than brunches, but in this model that's not true—and out in the world it might not be either.

(No doubt epidemiologists and those involved in contact tracing know this probably well, and they probably have some clever technique to deal with it.)

The code for this simulation model is on github.

Update: November 25, 2020

The *Times* contains multitudes so of course today it has an article on how many states are effectively giving up on much contact tracing. Amongst other examples:

> In North Dakota, state officials said last month that they could no longer have one-on-one conversations with everyone who may have been exposed. Aside from situations involving schools and health care facilities, people who test positive were advised to notify their own contacts, leaving residents largely on their own to follow the trail of the outbreak.

Needless to say, if you only actually contact trace cases in schools and health care facilities, you're going to discover that all your traceable cases come from schools and health care facilities…

Update: January 4, 2021

Rohan Alexander at the University of Toronoto used this post as a prompt for an undergraduate stats assignment, and one of his students, Annie Collins, built on the idea in her response.

Add comment

Comments are moderated and will not appear immediately.

Name

E-mail

Website

Message

Submit comment

Posted 2020-11-24. Last edited 2021-01-04.