Evaluating undercounts in epidemics

Michael Li, Jonathan Dushoff, David J. D. Earn, and Ben Bolker

05 September 2022

Introduction

Two papers (Böhning et al. (2020) and Maruotti et al. (2022)) have promoted a formula that claims to estimate the completeness of sampling of an infectious disease epidemic, using only case reporting counts. We believe this formula is fundamentally wrong, and that it is impossible to estimate undercounting without some kind of auxiliary information.

From Böhning et al. (2020):

> The idea is to apply this estimator (4) day-wise. We take an arbitrary day t. At this day we have $\Delta N(t)$ new infections. This will be viewed as f_1, the infected people identified just once. If we look at $\Delta N(t-1)$, then this is the count of new infections the day before. But these will still be infected at day t unless they decease. So, f_2 corresponds to $\Delta N(t-1) - \Delta D(t)$. We can ignore the number of recoveries as we are looking at infections which are very recent (notified at day t or $t-1$). Hence we are able to give the estimate for the number of hidden infections at day t as $H(t) = \frac{[\Delta N(t)]^2}{\Delta N(t-1) - \Delta D(t)}$

the authors go on to use a bias-corrected estimator from Chao.

Or from Maruotti et al. (2022):

> We will denote with $N(t)$ the cumulative count of infections at week t where $t = t_0, \ldots, t_m$. Hence $\Delta N(t) = N(t) - N(t-1)$ are the number of new infections at week t where $t = t_0, t_1, \ldots t_m \ldots$ Here $\Delta N(t)$ corresponds to the infected people identified just once, and $\Delta N(t-1)$ is the number of those identified twice.

In this paper the full bias-corrected expression (without a death term) (eq 1) is given as

$$\frac{\Delta N(t)(\Delta N(t)-1)}{1+\Delta N(t-1)};$$

the additional bias-correction (± 1) terms don't change our basic critique of the method.

Critique

This logic doesn't make sense. Why are cases identified at time $t-1$ representative of the number of cases counted twice? The fact that the same individual *could* be counted twice in the cumulative case report (for some sampling designs) is not relevant. How can comparing yesterday's count to today's provide information about the completeness of sampling, or anything other than the rate of growth of the epidemic?

In principle, it is possible to estimate "observability" or "catchability" if individuals are marked so they can be re-identified, or even with unmarked individuals given an appropriate sampling design (Royle and Dorazio 2008). However, public health case reporting is typically unrelated to mark-recapture (although it can be used to estimate completeness of contact tracing: Lerdsuwansri et al. (2022), Polonsky et al. (2022)). Furthermore, in practice case reports are often exclusive (i.e. someone who has been identified as a case will

not be reported again later), *or* are anonymous so that we cannot identify whether a particular case/infected individual was double-counted or not.

Exponential example

Suppose we have an epidemic that is growing exponentially. In the epidemic phase, the various measures of current epidemic size (incidence, prevalence, cumulative incidence, etc.) all grow exponentially, or geometrically if we sample in discrete time. Suppose $I(t)$ ("new cases" or incidence) is growing at a geometric rate λ per time step, i.e. $I(t) = I(0)\lambda^t$, and suppose we are sampling a fraction a (for "ascertainment") of the incidence. Again for simplicity, we'll see how the simple/non-bias-corrected estimate, without death or recovery, works out:

$$\begin{aligned} H(t) &= \frac{[\Delta N(t)]^2}{\Delta N(t-1)} \\ &= \frac{a^2 I(0)^2 \lambda^{2t}}{aI(0)\lambda^{t-1}} \\ &= aI(0)\lambda^{t+1} \\ &= a\lambda I(t) \end{aligned}$$

This shows that the number of hidden counts will be estimated to depend on the growth rate λ as well as the ascertainment ratio a — not what we wanted.

More generally, looking at the form of the expression $[\Delta N(t)]^2/\Delta N(t-1)$, we can see that the undercount will *always* be taken equal to the rate of acceleration of observed cases ($\Delta N(t)/\Delta N(t-1)$) times the number of new cases $\Delta N(t)$, so undercounting will always be assumed to be large (proportional to observed cases) at the beginning of the epidemic and continuously decreasing toward the end.

Simulation example (SIR model)

Simulating a simple SIR epidemic with a constant ascertainment rate 80% (i.e., an underreporting level of $(1-a)/a \cdot 100 = 25\%$) and applying the formula (Figure 1) incorrectly concludes that cases are underreported by an average of 92% (minimum, 13%; maximum 113%).

fixme (Mike) simplify model as suggested here so there are no hidden/undescribed complexities?

We conclude that the authors' suggested formula is invalid.

References

Böhning, Dankmar, Irene Rocchetti, Antonello Maruotti, and Heinz Holling. 2020. "Estimating the Undetected Infections in the Covid-19 Outbreak by Harnessing Capture–Recapture Methods." *International Journal of Infectious Diseases* 97 (August): 197–201. https://doi.org/10.1016/j.ijid.2020.06.009.

Lerdsuwansri, R., P. Sangnawakij, D. Böhning, C. Sansilapin, W. Chaifoo, Jonathan A. Polonsky, and Victor J. Del Rio Vilas. 2022. "Sensitivity of Contact-Tracing for COVID-19 in Thailand: A Capture-Recapture Application." *BMC Infectious Diseases* 22 (1): 101. https://doi.org/10.1186/s12879-022-07046-6.

Maruotti, Antonello, Dankmar Böhning, Irene Rocchetti, and Massimo Ciccozzi. 2022. "Estimating the Undetected Infections in the Monkeypox Outbreak." *Journal of Medical Virology* n/a (n/a). https://doi.org/10.1002/jmv.28099.

Polonsky, J., D. Böhning, M. Keita, S. Ahuka-Mundeke, J. Nsio-Mbeta, M. Mossoko, L. Kaiser, et al. 2022. "Novel Application of Capture-Recapture Methods to Estimate the Completeness of Contact Tracing During a Large Outbreak of Ebola Virus Disease, Democratic Republic of Congo, 2018-2020." *International Journal of Infectious Diseases*, Abstracts from the Eighth International Meeting on Emerging Diseases and Surveillance, IMED 2021, 116 (March): S98. https://doi.org/10.1016/j.ijid.2021.12.230.

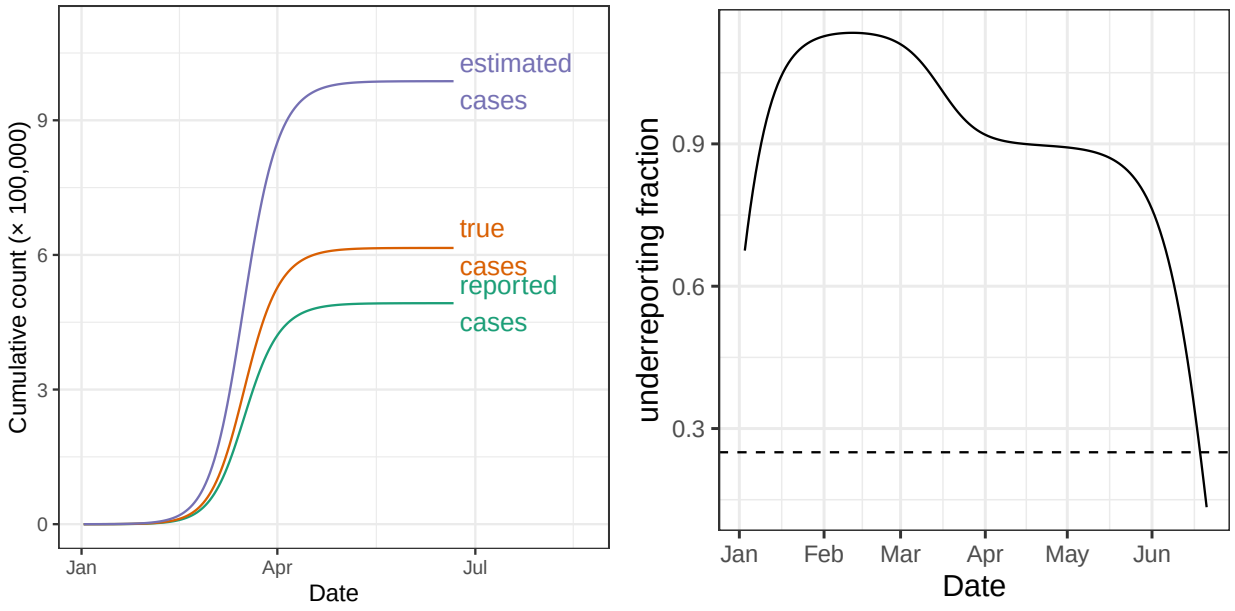

Figure 1: Simulated SIR epidemic with $\mathcal{R}_0 = 1.33$, infectious period of 3.33 days, $N = 10^6$, and ascertainment rate of 80% (undercounting fraction 0.25%).

Royle, J. Andrew, and Robert M. Dorazio. 2008. *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities.* Academic Press.