Evaluating undercounts in epidemics

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

07 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, based on a mark-recapture formula developed by Chao. We believe this formula is fundamentally wrong, and that it is impossible to estimate undercounting without a specialized sampling design or some kind of auxiliary information.

From Böhning et al. (2020):

> The idea is to apply [the Chao estimator] 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 [the number of infections counted twice] 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 $H(t) = \frac{[\Delta N(t)]^2}{\Delta N(t-1) - \Delta D(t)}$ …

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 $t = t_0, t_1, \ldots t_m$ … Here $\Delta N(t)$ corresponds to the infected people identified just once, and $\Delta N(t-1)$ is the number of those identified twice.

This latter paper does not use a death term, and gives the full bias-corrected expression (their eq 1) as

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

Critique

This approach is not an appropriate application of these mark-recapture formulas. 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 the number of unobserved individuals (hidden cases) if individuals can be re-identified when observed a second time, or even with unmarked individuals given an appropriate sampling design (Royle and Dorazio 2008). In practice public health case reporting rarely uses such sampling designs. Case reporting is usually exclusive (i.e. someone who has been identified as a case will not be reported again later) and/or anonymized so that we cannot identify whether a particular case/infected individual was double-counted or not. Mark-recapture methods are sometimes used in public health, but "one needs at least two sources of information with individual case reporting and a unique personal identifier for each case" (Desenclos and Hubert 1994).

Exponential example

Suppose an epidemic 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. For simplicity, we use the simple (non-bias-corrected) estimate, without death or recovery.

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

This shows that the estimated number of hidden counts depends on the growth rate λ as well as the ascertainment fraction a — not what we wanted. The "underreporting ratio" $H(t)/\Delta N(t) = H(t)/(aI(t)) = \lambda$ is independent of the ascertainment fraction. The bias-corrected formula, which we use in our simulations below, has similar problems.

More generally, looking at the form of the expression $[\Delta N(t)]^2/\Delta N(t-1)$, we can see that the estimate of underreporting will always be 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 underreporting will be assumed to be larger (proportional to observed cases) when disease incidence is increasing, and smaller when incidence is decreasing.

Simulation example (SIR model)

We simulate a deterministic SIR epidemic in discrete time, apply two different ascertainment fractions to derive a time series of case reports, and apply the authors' formula (including bias corrections) to estimate the number of hidden cases (Figure 1). At the peak of the epidemic, when the number of cases is large, the method predicts the same ratio of hidden cases to observed case reports regardless of the ascertainment fraction; at the beginning and end of the epidemic, the bias correction terms make underreporting appear lower in the lower-ascertainment scenario, when the number of reported cases is lower. In the high-ascertainment case ($a = 0.8$) the true ratio of hidden to reported cases is 0.25, but the average estimated ratio is 0.904 (range $0.037 - 1.093$); in the low-ascertainment case ($a = 0.2$) the true ratio is 4 and the mean estimate is 0.847 (range $0.004 - 1.085$).

We conclude that the authors' suggested formula should not be applied to disease outbreak incidence data.

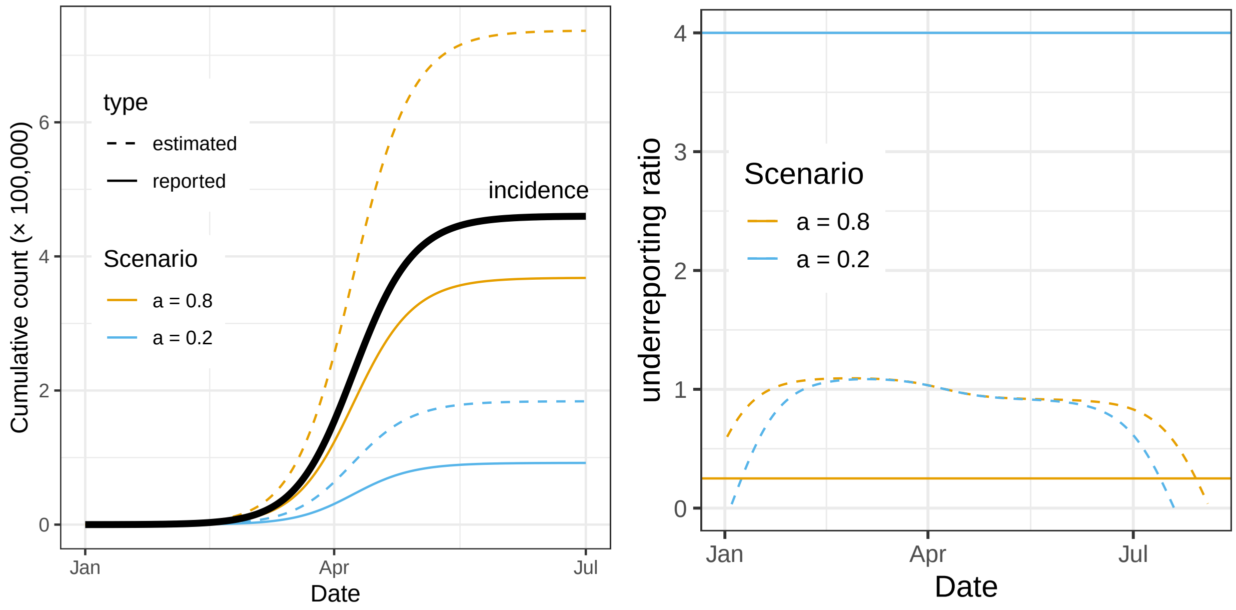

Figure 1: Simulated SIR epidemic with $\mathcal{R}_0 = 1.3$, infectious period of 3.33 days, $N = 10^6$. Left panel, cumulative numbers of reported and estimated-hidden cases along with true incidence (only one line appears because both scenarios have the same true incidence). Right panel, ratio of estimated (dashed) or true (solid) hidden cases to reported cases for each scenario.

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.

Desenclos, Jean-Claude, and Bruno Hubert. 1994. "Limitations to the Universal Use of Capture-Recapture Methods." *International Journal of Epidemiology* 23 (6): 1322–3. https://doi.org/10.1093/ije/23.6.1322.

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.

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.