Using Bayesian analysis to account for uncertainty and adjust for bias in coronavirus sampling[1]

Andrew Gelman and Bob Carpenter

7 June 2020

In early April 2020, a team of researchers based at Stanford University conducted an opt-in survey in the surrounding county, testing for coronavirus antibodies. The result was that 50 out of 3330 people in the survey—1.5%—tested positive. Extrapolating this to the population of the county as a whole yields an estimate of 29,000 exposed, which was much larger than the number of confirmed positive cases in the county (under 1000 at the time). Coronavirus tests were hard to come by at that time, and everyone knew that the number of confirmed cases was much less than the total number of people exposed, but it was not clear how much lower.

The Stanford study was posted on the preprint server medRxiv on April 11, and its authors were writing op-eds and explaining the implications of their findings on national television. The key result from their preprint: "a range between 48,000 and 81,000 people infected in Santa Clara County by early April, 50-85-fold more than the number of confirmed cases."

Three statistical questions arose:
1. Can we trust the results, given that the survey was not a random sample?
2. How did the raw rate of 1.5% in the data become an estimate of 2.5% to 4.2% in the preprint?
3. Where did the range of uncertainty come from, and is it appropriate given sampling variability in the data?

Questions 1 and 2 go together: the increase from 1.5% to 2.5% or more comes from a statistical adjustment done by the authors to correct for the sample not matching the population (as summarized by census totals for the county) by sex, ethnicity, and zip code. Unfortunately there are a few reasons we do not feel comfortable with these adjustments: first, they don't adjust for age; second, the adjustment for zip code is potentially very noisy (there are 58 zip codes in the county, which makes adjustment difficult, given that the sample contains only 50 positive tests); third, there is concern that, even after demographic and geographic adjustment, people who were more at risk were more likely to get tested; and, fourth, there are many "researcher degrees of freedom" in the adjustment process, leading us to be skeptical of any particular published result.

Question 3 is more challenging than it might seem at first, given that any estimate of prevalence must account for the specificity and sensitivity of the test—specificity is the

[1] To appear in the International Society for Bayesian Analysis Bulletin. We thank the U.S. National Science Foundation, National Institutes of Health, Institute for Education Sciences, and Office of Naval Research for support.

probability of getting a positive test, conditional on the true underlying state being positive, and the sensitivity is the probability of getting a negative test, conditional on the true underlying state being negative. But the specificity and sensitivity are not precisely known; they are estimated based on results from testing known positive and negative blood samples. Beyond this, the specificity and sensitivity can vary according to testing conditions.

During the week after the Stanford study appeared, there was increasing concern on social media regarding its data collection and statistical analysis, and it became clear that the calculations of confidence intervals in the preprint were wrong, even setting aside concerns about the demographic and geographic adjustments. In retrospect, it was not so easy to use classical statistical methods to account for all these uncertainties and adjustments at once.

But this is exactly the sort of problem where Bayesian analysis excels: combining information and propagating uncertainty from multiple data sources. Indeed, we were quickly able to program up a model in Stan to analyze the testing data more appropriately. Actually, we programmed up a series of models, starting with a simple analysis with uncertain specificity and sensitivity, then allowing the properties of the test to vary between experiments, then adding multilevel regression and poststratification (MRP) to adjust for measured differences between sample and population.

Based on our analysis, we do not think the data support the claim that the number of infections in Santa Clara County was between 50 and 85 times the count of cases reported at the time. These numbers are consistent with the data, but the data are also consistent with a near-zero infection rate in the county. The data in the study do not provide strong evidence about the number of people infected or the infection fatality ratio; the number of positive tests in the data is just too small, given uncertainty in the specificity of the test.

Unfortunately the Stanford team was not able to share their raw data with us, so we were not able to perform the MRP adjustment. However, our code is freely available, so they can perform this analysis with their data on their own computers.

Going forward, the analyses in this article suggest that future studies should be conducted with full awareness of the challenges of measuring specificity and sensitivity, that relevant variables be collected on study participants to facilitate inference for the general population, and that (de-identified) data be made accessible to external researchers.

Our paper describing our models and analyses is on medRxiv (https://www.medrxiv.org/content/10.1101/2020.05.22.20108944v2), and R and Stan code for the computations in the paper are on Github (https://bob-carpenter.github.io/diagnostic-testing/). In addition to explaining our models and fitting them to data, we also discuss informative hyperprior distributions for the hierarchical model (these are necessary because of the small number of experiments measuring specificity and sensitivity) and the challenge of summarizing posterior inferences near a boundary (in this case, the boundary of zero prevalence, which we know is not possible but is consistent with the data in this experiment).

We do not claim that Bayesian analysis was necessary to solve this problem. As with any statistical analysis, alternative approaches are possible that would use the same information and give similar results. But we will say that Bayesian inference for this example was transparent, direct, and relatively easy compared to the messy classical approximations used in the Stanford preprint. We hope that our paper and code can be a useful resource for future disease prevalence studies, as well as a jumping-off point for more elaborate models for more complex data including multiple tests, symptom reports, and additional patient-level information.